Unclassified
Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author)
Center for Naval Analyses of the Franklin Institute, Institute of Naval Studies

2a. REPORT SECURITY CLASSIFICATION
Unclassified

2b. GROUP

3 REPORT TITLE
PREDICTION OF REENLISTMENT USING REGRESSION ESTIMATION OF EVENT PROBABILITIES (REEP)

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)
Final

5 AUTHOR(S) (Last name, first name, initial)
Bryan, Joseph G. and Singer, Arnold

6 REPORT DATE
1 October 1965

7a. TOTAL NO. OF PAGES
97

7b. NO. OF REFS
23

8a. CONTRACT OR GRANT NO.
NONR 3732(00)

b. PROJECT NO.

c.

d.

9a. ORIGINATOR'S REPORT NUMBER(S)
Research Contribution No. 13

9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
INS U-65-11 195

10 AVAILABILITY/LIMITATION NOTICES All distribution of this report is controlled. Government agencies shall submit request to the Office of the Chief of Naval Operations. Other qualified users shall request through their sponsoring government agency to the Office of the Chief of Naval Operations.

11 SUPPLEMENTARY NOTES

12. SPONSORING MILITARY ACTIVITY
Office of Naval Research, Navy Department, Washington 25, D. C.

13. ABSTRACT

This report describes the statistical methodology of a "package" of computer programs referred to as REEP (Regression Estimation of Event Probabilities). REEP uses regression analysis techniques to arrive at equations which yield probabilities of occurrence for each of a set of possible events . An application of REEP to predicting reenlistment for enlisted Navy men is given.

Unclassified
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Regression analysis						
Prediction of reenlistment						
Retention of enlisted Navy men						
Dummy variables						
Stepwise screening						
Multiple regression						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2*a*. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2*b*. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7*a*. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7*b*. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8*a*. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8*b*, 8*c*, & 8*d*. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9*a*. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9*b*. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statements such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DDC users shall request through
____."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through
____."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through
____."

If the report has been furnished to the Office of Technical Services, Department of Commerce, for sale to the public, indicate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional explanatory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name of the departmental project office or laboratory sponsoring *(paying for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and factual summary of the document indicative of the report, even though it may also appear elsewhere in the body of the technical report. If additional space is required, a continuation sheet shall be attached.

It is highly desirable that the abstract of classified reports be unclassified. Each paragraph of the abstract shall end with an indication of the military security classification of the information in the paragraph, represented as *(TS)*, *(S)*, *(C)*, or *(U)*.

There is no limitation on the length of the abstract. However, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful terms or short phrases that characterize a report and may be used as index entries for cataloging the report. Key words must be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location, may be used as key words but will be followed by an indication of technical context. The assignment of links, rules, and weights is optional.

PREDICTION OF REENLISTMENT USING REGRESSION ESTIMATION OF EVENT PROBABILITIES (REEP)

Joseph G. Bryan
Arnold Singer

1 October 1965

INS Research Contribution No. 13

RESEARCH CONTRIBUTION
INSTITUTE OF NAVAL STUDIES

THIS RESEARCH CONTRIBUTION DOES NOT NECESSARILY REPRESENT THE VIEWS OF CNA OR U. S. NAVY. IT MAY BE MODIFIED OR WITHDRAWN AT ANY TIME.

REPRODUCTION IN WHOLE OR IN PART IS PERMITTED FOR ANY PURPOSE OF THE UNITED STATES GOVERNMENT.

CNA Center for Naval Analyses
WASHINGTON 25, D.C.

LR

THE INSTITUTE OF NAVAL STUDIES
OF THE CENTER FOR NAVAL ANALYSES

The Institute of Naval Studies engages in long-range studies of concern to the future Navy. Its basic mission is to assist the Chief of Naval Operations in establishing the long-range requirements of Naval operating forces for weapons, equipment, techniques, material, personnel, and supporting forces.

INS, the Operations Evaluation Group, and the Naval Warfare Analysis Group are the operating divisions of the Center for Naval Analyses. CNA, both through its operating divisions and through its own program of research in basic methodology and analysis techniques, conducts operations and systems research for the Chief of Naval Operations, the Commandant of the Marine Corps, and certain fleet and force commanders. CNA provides advice on operational problems susceptible of quantitative analysis, including the evaluation of new weapons, operational techniques, tactics, formulations of new requirements, technical aspects of strategic planning, and correlation of research and development programs with Navy and Marine Corps needs.

INS publishes three principal types of reports of its research, in addition to many memoranda of limited distribution:

STUDY, a complete, self-substantiating analysis, providing the Navy with a quantitative basis for executive decisions or for recommendations to higher authority. A study is endorsed by CNA or the operating division releasing the study, and represents the point of view held at the time of issue.

SUMMARY REPORT, a resume of research originally published in another form. A Summary Report presents results only, without substantiating analysis, and is designed for general information.

RESEARCH CONTRIBUTION, a paper of interest to workers in the field of operations research or the author's field.

Copies of INS publications may be requested from:

Director
Institute of Naval Studies
Center for Naval Analyses
545 Technology Square
Cambridge, Mass. 02139

DEPARTMENT OF THE NAVY
OFFICE OF THE CHIEF OF NAVAL OPERATIONS
WASHINGTON 25, D. C.

IN REPLY REFER TO
Op-911/res
Ser 509P91
3 NOV 1965

From: Chief of Naval Operations
To: Distribution List

Subj: INS Research Contribution No. 13; promulgation of

Encl: (1) INS Research Contribution No. 13, "Prediction of Reenlistment Using Regression Estimation of Event Probabilities (REEP)" by Joseph G. Bryan and Arnold Singer.

1. Enclosure (1) is a research contribution by the Institute of Naval Studies which was written in connection with a continuing program of studies concerning the problems of manning the future Navy and is based upon Annex A of the study "Manpower Considerations Applicable to the Navy in the 1970-80 Time Period". As a part of the latter study, the technique of regression estimation of event probabilities (REEP), as developed and packaged by the Travelers Research Center, was applied to the problem of reenlistment prediction for first term electronics ratings. The results show a high degree of correlation between the developmental and verification population samples, and provide significant discrimination between reenlistment probabilities within the sample group.

2. This paper is promulgated as a methodological example of the application of the REEP technique. This technique has been used previously in weather prediction and may have many future applications where predictions based on statistical analysis of past events are desired.

3. Appropriate Navy addressees who may desire to test the reenlistment prediction model within their own commands can request the identification of the predictors from the Director for Long Range Studies. The Director for Long Range Studies would be very interested in any significant confirmation of the prediction model. Comments and recommendations are invited.

4. Requests for additional copies of enclosure (1) should be sent to:

Director for Long Range Studies (Op-911)
545 Technology Square
Cambridge, Massachusetts 02139

J. M. LEE
By direction

Distribution:
See next page

Op-911/res
Ser 509P91

Distribution

SECNAV
CNR (3)
OP-07
OP-09
OP-090
OP-93
OP-01
OP-03
OP-911
USNA ANNA (2)
SUPTNAVPGSCOL (2)
PRESNAVWARCOL (2)
BUWEPS (3)
BUPERS
BUPERS (Pers Aa1)
NAVPERSRSCHACT, Wash.,D.C.
NAVPERSRSCHACT, SDiego

INS Research Contribution No. 13

Institute of Naval Studies
Center for Naval Analyses
The Franklin Institute

PREDICTION OF REENLISTMENT USING REGRESSION ESTIMATION OF EVENT PROBABILITIES (REEP)

by

Joseph G. Bryan*
and
Arnold Singer

Joseph G. Bryan

Arnold Singer

1 October 1965

This Research Contribution does not
necessarily represent the views of
CNA or the U. S. Navy. It may be
modified or withdrawn at any time.

Work conducted under Contract No. NONR-3732(00)

*The Travelers Research Center, Inc., Hartford, Connecticut

ABSTRACT

This report describes the statistical methodology of a "package" of computer programs referred to as REEP (Regression Estimation of Event Probabilities). REEP uses regression analysis techniques to arrive at equations which yield probabilities of occurrence for each of a set of possible events. An application of REEP to predicting reenlistment for enlisted Navy men is given.

SYNOPSIS

This Research Contribution describes a package of statistical programs dealing with regression analysis. The application of these programs, as described herein, is addressed to the following problem:

> Using available records on eligible first term electronics men (e.g., records such as age, education, test scores, length of military duty, recruiting area) what sort of index or statistical digest of this information can be devised so as best to distinghish evential reenlistees from non-reenlistees?

A well tested methodology which goes by the acronym REEP (Regression Estimation of Event Probabilities) had already been developed for this type of problem. This report describes this methodology in two styles of writing. In one style, used in section II as an introduction, the main ideas are presented in language that does not presuppose specialization in statistics. In the other, technical style, an exposition intended for practicing statisticians is presented (see sections III through V). The general methodological problem is to identify the independent components of information and then to combine these components in a single formula. The discussion of methodology is followed by a detailed account of its specific application to the retention of eligible first term electronics men using data which pertained to those men who took reenlistment action at some time between August 1962 and July 1963, inclusive, less the month of October 1962.

REEP uses a mathematical model to relate the probability of

the occurrence of a designated event referred to as the predictand (in the present application, reenlistment) to a group of independent predictor variables (for example, age, education, state of residence) chosen as statistical indicators of the event. By dividing the independent variables into discrete classes and associating a dummy variable (a zero-one index) with each class, reenlistment rate curves of arbitrary shape can be approximated. Using dummy variables also enables one to describe the response to qualitative as well as quantitative variables. The combined effect of many variables is represented by an expansion in terms of these dummy variables; the latter can be associated with compound classes pertaining to combinations of two or more variables, as well as with simple classes pertaining to individual variables. A screening process is used to eliminate redundancy, and the coefficients (measuring the net effects) of the retained dummy variables are determined by least squares. All fitting of data is done on a developmental sample. The final results are checked on an independent verification sample, and this check furnishes an estimate of performance if the same formula were to be used on still other data.

Two types of index, differing in the mathematical terms employed, were derived and tested. The first index, referred to as Model A, was made up of dummy variables which pertained only to univariates, i.e., terms that individually involved only one variable at a time. For example, one such term had to do with the individual's age, and another, with the Recruiting Area from which he entered the Navy. The second index, referred to as Model B, was made up of a number of terms that individually could involve either just one variable at a time (univariates) or two at a time (bivariates) -- for example, a combination of an aptitude test score and educational level.

Upon testing the Model A index on an independent data sample, not used in any way for its derivation, it was found that this index enables the user to segregate first term electronics men, in advance of reenlistment action (and, in fact, just about at the time of recruitment into the Navy), into three groups:

(1) A group, comprising about 18% of the eligibles, for which the reenlistment rate (.188) is significantly lower than average;

(2) a group, comprising about 58% of the eligibles, for which the reenlistment rate (.248) is about average; and

(3) a group, comprising about 24% of the eligibles, for which the reenlistment rate (.322) is significantly higher than average.

Similarly, upon testing the Model B index on the same independent data sample used for testing the Model A index, but which was not used in any way to derive the Model B index, it was found that the Model B index enables the user to segregate first term electronics men, in advance of reenlistment action (and, again, just about at the time of recruitment into the Navy), also into three groups:

(1) A group, comprising about 27% of the eligibles, for which the reenlistment rate (.193) is significantly lower than average;

(2) a group, comprising about 44% of the eligibles, for which the reenlistment rate (.252) is about average; and

(3) a group, comprising about 29% of the eligibles, for which the reenlistment rate (.315) is significantly higher than average.

By a statistical test of significance, it was established that the differences of discrimination between the models were not due to chance. Therefore, it is concluded that Model B affords significantly sharper distinctions than does Model A.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF TABLES (Continued)

I. INTRODUCTION

The work being reported in this Research Contribution was performed as a part of the second phase of the Multivariate Study of Enlisted Retention (MUSTER)* -- a sub-study of the INS Manning Study. The object of the MUSTER study was to apply multivariate techniques of analysis to a large base of data to identify factors that are significantly related to reenlistment behavior of Navy men.

The analyses performed during the first phase of the MUSTER study concentrated on studying the effects on reenlistment of each of a number of individual variables. This was done to help the analyst to understand better the variables with which he is dealing, and to formulate hypotheses on the basis of the results of such a preliminary study. During the second phase of the MUSTER study the analyses were aimed at pinpointing interacting effects of variables and isolating significant relationships between variables (or combinations of variables) and reenlistment.

To accomplish this end, the research conducted during the second phase was undertaken using two different approaches. First, the type of contingency count analyses performed in the first phase was continued but at a deeper multivariate level. Reenlistment rates and other pertinent statistics were calculated as a function of one variable while holding one or more other variables fixed. This approach enables one to factor out the interacting effect of these additional variables on the first variable, and to concentrate on the relationship between the first variable and reenlistment. Many combinations of variables were tested in this manner for their joint effect on reenlistment.

*See references (17) and (18).

In addition to the analyses based on contingency counts, a second type of investigation was performed. Regression analysis techniques were applied* to determine which variables are most significantly related to reenlistment. These analyses were carried out using a "package" of statistical computer programs known as REEP (Regression Estimation of Event Probabilities).** REEP analyzed the characteristics of the men in the MUSTER population and, on the basis of these characteristics, developed an equation which enables one to predict the probability of reenlistment for other men.

This report contains a documentation of the REEP statistical methodology and the results of the application of that methodology to the problem of predicting reenlistment behavior.

Section II contains a <u>qualitative</u> discussion of the object and mechanics of the REEP program, while sections III through V offer a more detailed, <u>quantitative</u>, and analytical review of REEP. Sections VI and VII describe the results of the application of REEP to reenlistment prediction.

Many people have contributed in various ways to the development of REEP. The ideas of dummying the predictors and estimating probabilities by regression were proposed in 1955 by I. Lund (see references (8) and (9)). Related work has been published by D. R. Cox (references (3) and (4)) and by S. L. Warner.

*The application described later in sections VI and VII deal with a sample of first termers in electronics ratings.

**The Travelers Research Center, Inc. (TRC), Hartford, Connecticut, who had earlier developed the programs, assisted in this portion of the analysis.

(reference (23)). The present REEP package with its full complement of statistical procedures and high-speed computer programs was designed and engineered at TRC by R. G. Miller in collaboration with T. G. Johnson (reference (11)) as an outgrowth of their experience in the development of multiple discriminant analysis (reference (12)).

Important contributions in the development of the MUSTER study were made by A. S. Morton. Guidance and encouragement throughout the MUSTER study were provided by H. K. Gayer, director of the total effort of the INS Manning Study project. Computer programming support for this report was provided by T. G. Johnson of The Travelers Research Center, Inc., and C. R. Berndston of INS.

II. GENERAL DISCUSSION OF REEP

A. Object of REEP

REEP provides estimates of the probability that a designated event (the predictand) will occur under specified circumstances. With reference to the retention of enlisted personnel, the predictand is reenlistment, and the specified circumstances are described by stated combinations of available records (predictor variables) on individual men -- predictors such as age, education, scores on various tests, place of birth, recruiting area, etc. The separate indications of reenlistment potential from the different records on each man are weighed and balanced by means of a mathematical formula, called a regression* function, and the calculated resultant indication provides the estimated probability of reenlistment for that man.

REEP was developed as a practicable method of extracting the combined predictive content of many statistical predictors bearing on the same predictand. It is applicable not only to the estimation of the probability that a single designated event (such as reenlistment) will occur, but also to the broader problem of estimating the respective probabilities that any one of several possible events will occur. It is well suited to the needs of statistical decision making, since it furnishes the requisite probability distributions, conditional on observable antecedents. This general exposition, however, will be directed

* Through historical accident, the word "regression" has acquired the technical meaning of "pertaining to statistical prediction."

toward the immediate end of using REEP to estimate reenlistment rates. An appreciation for its rationale may be gained by reviewing some of the obstacles REEP was designed to overcome.

B. Regression Analysis vs. Contingency Counts Analysis

Data permitting, a conceptually simple way of assessing the combined effects of many simultaneous indicators would be to divide each indicator into a series of classes, construct compound classes from combinations of the simple classes (pertaining to the individual indicators), and for each compound class make a direct computation of reenlistment rate from a contingency count of the available data. For instance, age might be divided into a number of relevant levels, education might be classified by the number of years of schooling completed, and point scores on each test bracketed according to ranges that would make sense for the test in question. A compound class would be specified by stipulating the simultaneous levels of a group of characteristics, e.g., age between 18½ and 19 years, 15 years of education, General Classification Test score between 45 and 50, Arithmetic Test score between 50 and 55, etc. The reenlistment rate for such a compound class, in the sample considered, would be estimated by dividing the number of actual reenlistments by the total number of men in that compound class who were eligible for reenlistment.

Such analyses based on contingency counts are extremely useful when dealing with up to three or possibly four variables simultaneously. However, when one wishes to consider the joint effects of a large number of variables, a more feasible approach is to formulate a model that describes the mathematical dependency of the variables. By means of the model, the outcome corresponding to any possible combination of the variables can be estimated by computation from the indicated formula. REEP employs a model that has considerable flexibility in representing mathematical relationships by statistical approximations, since for the type of expansion used in REEP, the true mathematical relationship need not be known in advance. The mathematical terms in the expansion have adjustable coefficients that can be fitted to the data, thus yielding a final formula that is reconciled with the actual body of experience.

C. Dummy Variables

The variables used as predictors of a predictand may be quantitative (e.g., age, education, scores on tests) or qualitative (e.g., place of birth, recruiting area). The type of expansion used in REEP to evaluate the combined predictive information of simultaneous predictors was designed to apply equally well to both kinds, and thus to mixtures of quantitative and qualitative predictors. As suggested by the age-old device of reckoning the area under a curve with the aid of a series of rectangles, a useful approximation to a mathematical curve can be obtained with a step-like configuration, called a step function (see figure 1). The expansion used in the REEP equation is based on step functions. Besides offering certain worthwhile advantages* in representing curves with even the most complex non-linearities, step functions are ideally suited to the analysis of qualitative variables which would otherwise be difficult to handle in a mathematical model.

*Among these advantages are the boundedness of the approximations, the ease of retaining a sufficient number of significant figures to insure numerical accuracy in computations, the great economy of machine storage gained by binary variables, and the spectacular increase in computational speed made possible with binary variables by the fact that calculations of products make use of the highest speed "logical and" operations in place of the much slower multiplying operations.

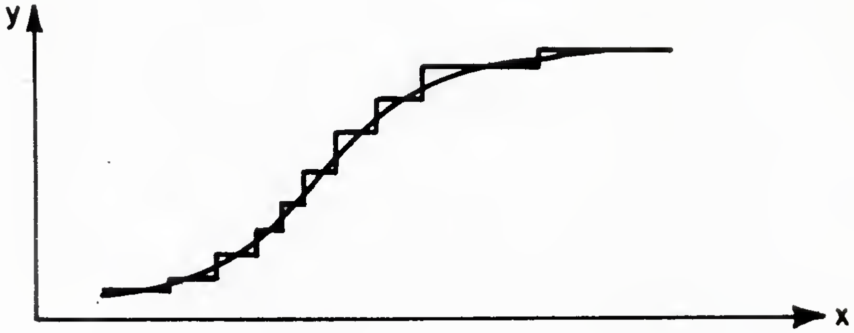

FIG. 1: APPROXIMATION OF A CURVE BY MEANS OF A STEP FUNCTION

Purely as a mathematical convenience (but a great one) in constructing step functions, it is helpful to divide the process into three parts: (1) dividing each independent variable (statistical predictor) into suitable classes, (2) associating (as will be explained) a separate dummy variable with each class of an independent variable, and (3) assigning a numerical coefficient to each dummy variable. The first operation (classification) defines the width of any step, and the third (assigning a coefficient) defines its total height "above the floor." The second operation (associating a separate dummy variable with each class) is a mathematical trick to simplify the bookkeeping required in the calculation of coefficients. A dummy variable is simply an index used to denote the occurrence or non-occurrence of a given characteristic, e.g., having 12 years of education. The index has only two admissible values, zero or one. When the given characteristic occurs, the index value is one; when the given characteristic does not occur, the index value is zero. Thus, dummy variables provide a convenient way of tagging data, by showing what characteristics are represented on each record.

D. Objective Dummying

Breaking up an independent variable into suitable classes can be done subjectively, if the user of REEP prefers to use his personal judgment, or it can be done objectively through the utilization of statistical analyses. The objective process is called objective dummying. Starting with a large number of fine subdivisions, the objective dummying process uses a statistical test* to decide whether or not the predictand (here, reenlistment rate) has a significantly different response to one class of an independent variable than to another class of the same variable. Classes with significantly different responses are kept separate; those with the same or not significantly different responses may be consolidated.

Objective dummying can be done in either of two fashions, an ordered fashion or an unordered fashion. In the ordered fashion, only adjacent classes of an independent variable can be combined. For example, if no significant differences are found one might wish to combine the classes corresponding to 10 years of education and 11 years. However, one might not wish to combine the men with 10 years together with those having 14 years of education -- even if it were found that their responses to reenlistment are not significantly different. In the unordered fashion, similarity of response shown by the predictand is the

*The test employed is the Kolmogorov-Smirnov two-sided test for two samples. See section IV.

only condition for combining classes of the independent variable. Typically, an ordered grouping is chosen for a quantitative independent variable, and an unordered grouping for a qualitative variable. Once the classes have been decided upon, the dummy variables are entered on the data tape by applying a special processing program called a dummy-generation program.

E. Superposition of Predictive Components

The next question to consider is how to go about assigning appropriate coefficients to the dummy variables. The possibility of representing a single curve by a step function has been mentioned, but nothing has been said yet about combining the indications of two or more independent variables. The REEP model is one of superposition. The superposition principle is the time-honored basis of many engineering applications of mathematics. It amounts to an assumption that the resultant of any number of influences is the algebraic sum of their separate components. If the net effect* of class X of variable A is to elevate the reenlistment rate by 15 percentage points above the average, and the net effect of class Y of variable B is to elevate the reenlistment rate by 10 percentage points above the average, where these net effects are estimated under the assumption of superposition, then the combined effect of these two classes (X of variable A, reinforced by Y of variable B) would be to elevate reenlistment rate by 25 percentage points. In the REEP model, the superposition principle is

*The net effect of a given class of a stated predictor is the residual predictive component ascribable to it, after allowance has been made for the simultaneous effects of the other predictors. In a linear function of dummy variables, such allowance is made automatically, when coefficients are determined simultaneously by least squares.

expressed by a linear function of dummy-variable predictors. The assignment of coefficients to the dummy variables is done by the method of least squares.* Its aim may be interpreted as one of determining net effects of the respective classes that characterize the independent variables.

If the assumption of superposition seems restrictive, it should be remembered that the REEP model is expressed mathematically by an expansion, and that the terms of the expansion can be made as general as the user is prepared to handle. Dummy variables can be extended to compound classes, and these can be made as elaborate as available data, funds, and computer facilities permit. In the present application, it was found feasible to construct compound classes from pairs of variables. Thus synergistic effects were investigated to the extent of examining pairwise interactions.

F. Screening Procedure

Redundancy of information is commonly experienced in statistical analysis. Experiments testing the accuracy of reconstructing a dependent variable by formulas using many independent variables have repeatedly brought out the fact that very nearly all of the relevant information is concentrated in a comparatively small number of the variables employed. Moreover, it is consistently found that when alternative formulas are applied to fresh data, the greater precision is demonstrated by the formulas using the smaller number of variables, where that smaller set is restricted to those variables previously shown

*This method minimizes the sum of the squares of the errors made in the individual predictions of reenlistment.

to contain most of the information. For this reason, a screening procedure is an integral part of REEP.

By the screening procedure, the variables are ranked preferentially, in order of their incremental contributions to the reconstruction of the predictand. The dummy variable that makes the greatest single contribution is ranked first; with this rank position fixed, the dummy variable that makes the greatest addition to the first contribution is ranked second; and in general, each rank in turn is filled by the dummy variable that offers the greatest increase in precision, when added to the set already chosen. When none of the remaining dummy variables adds significantly to the precision already attained, the screening terminates, and these remaining variables are discarded as redundant. This elimination of redundancy has a stabilizing effect on the determination of coefficients and improves the reliability of the prediction formula when applied to new data.

G. Developmental and Verification Samples

An accepted practice in connection with least squares is to compute estimated standard errors ascribed to the respective coefficients. Conventional formulas for standard errors of coefficients do not hold for screened predictors (see reference (1)), and even if correct formulas were available, error estimates for the separate coefficients would not equip the user to estimate the precision of the whole formula, when applied to new data. To provide such an estimate, the REEP program makes use of two samples. One sample, ordinarily the larger of the two, is termed the developmental sample and is used for all calculations that affect the objective dummying, the screening, and the determination of coefficients. The other sample, termed the verification sample, is kept segregated from the first sample and is used exclusively for testing purposes. The formula derived from the developmental sample is tested on the verification sample to obtain an independent estimate of the precision to be expected when the formula is applied to new data.

H. Summary

In summary, REEP uses a mathematical model to relate the probability of the occurrence of a designated event referred to as the predictand (in the present case, reenlistment) to a group of independent predictor variables (for example, age, education, state of residence) chosen as statistical indicators of the event. By dividing independent variables into discrete classes and associating a dummy variable (a zero-one index) with each class, reenlistment rate curves of any shape can be approximated. Using dummy variables also enables one to describe the response to qualitative as well as quantitative variables. The combined effect of many variables is represented by an expansion in terms of these dummy variables; the latter can be associated with compound classes pertaining to combinations of two or more variables, as well as with simple classes pertaining to individual variables. A screening process is used to eliminate redundancy, and the coefficients (measuring the net effects) of the retained dummy variables are determined by least squares. All fitting of data is done on a developmental sample. The final results are checked on an independent verification sample, and this check furnishes an estimate of performance if the same formula were to be used on still other data.

III. ANALYTICAL THEORY OF REEP

A. Introduction

A variable that is the object of prediction or estimation will be called the predictand (in the present application, reenlistment), and the variables used to arrive at the prediction or estimation will be called predictors (for example, age, education, test scores). The type of prediction problem under consideration is that in which the predictand can assume any one of several distinct values, levels, or states (in the application to reenlistment prediction there are two states, viz., reenlist or not); and the object is to make use of the information available in the predictors to estimate the respective probabilities associated with each possible predictand state -- that is, to estimate the chances that any specified state will be the one that the predictand actually assumes in a given instance.

Let the number of distinct states of the predictand be denoted by G. Unless otherwise noted, it will be taken for granted that these G states* are exhaustive (some one of them must necessarily occur) and mutually exclusive (no more than one state can occur at a time). If the predictors uniquely determined the predictand, the probability would be unity for some one predictand state, as fixed by the predictors, and zero for all others. In a real situation, however, the predictors merely influence the probability by tending to favor the occurrences of some states more than others, depending on the given values of

*Again, with reference to personnel retention, G=2, because a man can decide either for or against reenlistment.

the predictors, and the probability of occurrence is distributed over all G states. The statistical problem is to describe this distribution in terms of the predictors.

The REEP approach to this problem is by way of multiple regression analysis. An alternative approach, by way of discriminant analysis, is expounded in reference (12), and a comprehensive comparison of discriminant analysis and REEP is presented in reference (11). In performance, the two alternatives appear equally good, but practical advantages (speed and economy) are heavily on the side of REEP. In the regression approach, a dummy variable $D_g(g=1,2,\ldots, G)$ is associated with each state g of the predictand: $D_g=1$ if state g occurs, or $D_g=0$ if state g does not occur. Each dummy variable D_g, in turn, is treated as a predictand, to be estimated by a separate regression function (one for each dummy predictand). The device of using a common set of predictors for all $D_g(g=1,2,\ldots G)$, as REEP does, insures that the sum of the estimated probabilities will be identically equal to unity in every instance.

In the strict definition of the term, a regression function defines the conditional mean value of a predictand for any specified set of values of the predictors. Now the true conditional mean value of a dummy variable D_g is identically equal to the relative frequency -- hence, the conditional probability -- of the occurrence of state g, under the conditions defined by the predictors. If the exact mathematical specification of the regression function could be given, the true conditional probabilities could be determined from it.

In actuality, of course, the mathematical specification of the regression function is not available. As a serviceable approximation to that function, REEP employs a linear expansion in terms of dummy variables, constructed from the predictors.

These dummy variables can represent simple classes pertaining to individual variables, or, as desired, compound classes pertaining to combinations of two or more variables.

Hereafter, the term "regression function" will be applied to the expansion employed in REEP. The regression function $\hat{D}_g$ for D_g is of the form

$$\hat{D}_g = A_{0g} + A_{1g}X_1 + A_{2g}X_2 + \dots + A_{Mg}X_M \qquad (1)$$

The predictors $X_1, X_2, \dots, X_M$ are selected by screening (see next section), and the base constant A_{0g} and the coefficients $A_{1g}, \dots, A_{Mg}$ are determined by least squares, so as to minimize the average value of the squared discrepancy $(D_g - \hat{D}_g)^2$.

As a consequence of the simultaneous equations by which the A's are determined, it can be shown that the respective base constants $A_{01}, A_{02}, \dots, A_{0G}$ sum to unity, and the respective coefficients A_{j1} , $A_{j2}, \dots, A_{jG}$ of each predictor X_j sum to zero. Therefore the probability estimates $\hat{D}_g$ sum to unity for all possible predictor values:

$$\sum_{g=1}^{G} \hat{D}_g \equiv 1 \qquad (2)$$

A complete analytical proof of this fact is given in section III-E.

Another interesting sidelight on the simultaneous equations of least squares, in the present context, is the reconstruction of marginal relative frequencies. The simultaneous equations can be expressed as follows:

$$\sum_i \hat{D}_{gi} = \sum_i D_{gi}$$

$$\sum_i \hat{D}_{gi} X_{ji} = \sum_i D_{gi} X_{ji} \qquad (3)$$

where g is fixed, and the second subscript i pertains to an individual case. Since the predictors are dummy variables, these equations imply that the expected number of occurrences of state g, according to the formula, is made equal to the actual number of occurrences of state g observed in the sample in every class of every predictor. Hence, a respectable degree of correspondence with the actual data is guaranteed.

Because $\hat{D}_g$ is only an approximation to the true conditional mean value of D_g, it is possible for a formula of this type to yield an inadmissible estimate of probability, with certain predictor combinations. An estimate is inadmissible if it is either negative or greater than 1. Since $\sum_g \hat{D}_g \equiv 1$, a value of $\hat{D}_g$ in excess of unity is always accompanied by at least one negative value (for some other g), but not conversely, i.e., the presence of a negative value does not necessarily imply that one of the remaining values exceeds unity. If a value of $\hat{D}_g$ is inadmissible on a given occasion, the REEP program automatically replaces all of the G estimates on that occasion by revised estimates P_g ($g = 1, 2, \ldots, G$). The revised estimates are defined as:

$$P_g = \tilde{D}_g / \Sigma \tilde{D}_g, \qquad (4a)$$

where

$$\tilde{D}_g = \begin{cases} 0 & \text{if } \hat{D}_g \leq 0 \\ \hat{D}_G & \text{if } 0 < \hat{D}_g < 1 \\ 1 & \text{if } \hat{D}_g \geq 1 \end{cases} \tag{4b}$$

This rule would leave $\hat{D}_g$ unchanged if the initial estimates were all admissible.

B. Predictor Screening

The number of dummy variables generated by a set of ordinary variables will be equal to the sum of the numbers of classes into which the ordinary variables are divided. The number of independent dummy variables pertaining to a given ordinary variable will be one less than the number of classes, inasmuch as the classes are exhaustive and mutually exclusive. Even so, making allowance for this reduction in the count, a modest number of ordinary variables can lead to a sizable number of independent dummy variables. A screening procedure is used to select a manageable number of dummy predictors that account for most of the predictive information, without redundancy.

Let the initial set of tentative dummy predictors under consideration be designated as T_1, T_2, ..., T_Q. The screening involves the computation of variance-ratio statistics F, where an individual F tests the significance of an additional predictor. If R_k denotes the multiple correlation coefficient computed from k predictors, and d.f. stands for the estimated number of degrees of freedom left at stage k, the variance ratio F used to test the k-th selection is given by the equation

$$F = \frac{R_k^2 - R_{k-1}^2}{1 - R_k^2} \times (\text{d.f.}) \qquad (5)$$

Assuming independent observations (a justifiable assumption with reference to reenlistment decisions), the value of d.f. for the k-th selection is N-k-1, where N is the sample size. Had serial correlation been present (as happens, for example, with weather data), then N would have to be reduced somewhat. In the REEP program, this reduction when required, is made on the basis of the "Runs Test" (see reference (6), p. 12).

The screening takes account of all G predictand states simultaneously and is done as follows. Compute a value of F for each tentative predictor T_q ($q = 1, 2, \ldots, Q$) in relation to each dummy predictand D_g ($g = 1, 2, \ldots, G$). At the first stage of predictor selection, there will be G x Q values of F (since G values will be obtained for each T_q). Denote by X_1 the predictor that yields the largest single value of F out of all of these G x Q values. This predictor X_1 is called the first predictor.

The screening process is now repeated to select a second predictor. For each D_g, trial multiple correlations using two predictors are computed. The two predictors on any trial are X_1 and one of the remaining T's. There will be G(Q - 1) such multiple correlations and so there will be the same number of F-values. The trial predictor yielding the largest value of F among these G(Q - 1) values is selected as the second predictor and is denoted by X_2.

The screening is continued to select third, fourth, and further predictors until M predictors $X_1, X_2, \ldots, X_M$ have been chosen. As each predictor is selected, a statistical test comparing the highest computed F-value with a certain critical value of F is employed to decide whether the proposed, selected predictor appears to be useful. The termination point M is established by the fact that χ_M passes this test, but the next candidate (which, if successful, would be called χ_{M+1}) fails it.

The rule for setting the critical level of F, used for terminating the screening process, was suggested, in part, by the theory of ordered statistics. If the distribution function of a continuous variate x is $P(x)$, then the distribution function $\Phi_n(\xi)$ of the maximum value ξ among n independent observations of x is given by

$$\Phi_n(\xi) = [P(\xi)]^n \tag{6}$$

(See reference (7) p. 75.) If α is a chosen significance level for deciding whether or not a predictor should be judged useful, the critical level F_c of F is determined (essentially) by solving the equation

$$[P(F_c)]^n = 1 - \alpha \tag{7}$$

where $P(F)$ denotes the distribution function of F. Although it would be perfectly possible to solve equation (7) by logarithms and table look-up, the REEP program substitutes an approximation. Setting $p_c = 1 - P(F_c)$, expanding $(1 - p_c)^n$ as a binomial, and retaining only the first-degree term in p_c, we obtain

$$p_c = \frac{\alpha}{n} \tag{8}$$

By use of a normalizing transformation of F, since p_c is usually beyond the range of F-tables, the value F_c of F is determined as that for which $1 - P(F_c) = p_c$. This rule would be appropriate if the computed F-values were independent. Since the predictors are correlated, however, the test may be overly strict. As a compromise between the facts that (1) the predictors are not independent, (2) G trials are made on each predictor tested at stage k, but (3) these G trials are definitely not independent, the number n_k used for n at stage k is taken as

$$n_k = Q - (k - 1) \tag{9}$$

Certainly no claim for rigor can be made for this formula, but according to experience so far, the test does not seem to err on the side of laxness, and indeed may still leave us with an overly strict criterion of significance.

As the REEP program stands at present, the maximum value of G is 10, that of Q is 500, and that of M is 36. Although there is no theoretical limit that one can place on M, it has been found in practice that the screening regularly cuts off much before M = 36. This choice of limit on M, however, is based solely on experienced judgment.

The process just described is called forward screening to distinguish it from a different, but related, selection process, called backward screening. In backward screening, a definite set of B ($B \leq Q$) trial predictors is chosen to begin with, and a regression formula based on all B predictors is determined. The least important predictor is then identified by calculating the increase in mean square error due to the omission of each predictor, in turn, when the other B - 1 predictors are retained. If the least important predictor is judged non-significant, it is eliminated. The tests are applied again to the remaining set of B - 1 predictors, and the deletion process is continued in a stepwise fashion, analogous to that used in forward screening.

Because forward screening can cope with a much larger set of tentative predictors than can backward screening (to the best of the writers' knowledge of feasible computational capacities, B falls far short of 500) forward screening has been chosen for REEP.

With stepwise screening, there exists the possibility of overlooking some peculiarly potent subset of available predictors. The risk of this oversight has troubled many investigators and has inspired attempts to explore all possible combinations of k out of Q tentative predictors. Such an undertaking is flatly out of the question, unless Q and k are both comparatively small. With Q = 500 and merely with k = 3, there can be some millions of combinations to try. Furthermore, the questions of independence and statistical significance raised earlier in this subsection become vastly more complicated.

A mathematical study of predictor selection by J. Oosterhoof (reference (16)) concludes with the following statement:

> "Reviewing the results, we see that a class of examples can be constructed where forward selection and backward elimination do not lead to optimal k-subsets, even if both methods yield identical sequences of the independent variables. The k-subset they produce can be a bad one in a quantitative sense, that is, there are many better k-subsets, as well as in a qualitative sense, that is, there exists at least one k-subset that is very much better. Furthermore, it is possible that both methods, though identical, do not lead to optimal k-subsets for any k except k=1 and k=m-1. In some cases it is possible to detect such anomalies by inspection of the correlation matrix: a highly intercorrelated subset of independent variables which appear in the regression equation only at a later stage, may be a sign of misbehavior.
>
> "However, in our opinion a better and yet not too troublesome method will be hard to find, because such a method should essentially use the correlations between all variables at every stage of the process."

C. Use of Developmental and Verification Samples

The ordinary formulas for estimating the sampling variability of regression constants do not hold when the predictors are required to meet preliminary tests of significance, as they are in selective screening (see reference (1)). The most important single question to answer is not that of the sampling behavior of separate coefficients, but rather that of the sampling behavior of the estimated regression function as a whole. In REEP, the latter question is attacked by reserving an independent sample for purposes of verification.

By a randomization technique, the initial sample is divided into two parts. One part, usually the larger, is called the developmental sample and is used for all processes involving accommodation to the predictand -- objective dummying, predictor screening, fitting of constants. The other part, called the verification sample, is used solely to obtain estimates of predictive accuracy when the regression formulas are applied to independent data. The program can accept a developmental sample size of about 10,000 and, if desired, an even larger sample size for verification.

D. Performance Measures

Predictive performance is measured by the correspondence between P_g and D_g in the verification sample. (As noted in section III-A, P_g reduces to $\hat{D}_g$ when no adjustment is required.) Although the verification sample gives the proof of the pudding, performance in the developmental sample is also put on record. Several kinds of evidence are presented on the computer printout (not all of which will be described here).

An overall measure of correspondence between P_g and D_g is given by the mean-square error, as defined by the Brier P-Score. For a single probability forecast of G states, the

P-score is defined as

$$\text{P-score} = \sum_{g=1}^{G} (P_g - D_g)^2 \qquad (10)$$

A P-score of 0.0 indicates a perfect forecast; the poorest score is 2.0, which results when for some value g, $P_g = 1$, whereas in fact there exists some other value g' such that $D_{g'} = 1$. In comparing two forecasts of the same events, the lower P-score indicates the better forecast. For a series of N probability forecasts of G states, the P-score is defined as follows:

$$\text{P-score} = \frac{1}{N} \sum_{i=1}^{N} \sum_{g=1}^{G} (P_{gi} - D_{gi})^2 \qquad (11)$$

Tabular comparisons are made between estimated and actual frequencies of occurrence of each state g. The REEP probability forecasts P_{gi} are sorted into 12 classes using the class limits listed in column 2 of table 1. For each state g of the predictand and for each of the 12 classes of P_{gi} separately, four summary statistics are computed and printed: a count of the number of times (F) that the value of P_{gi} falls within the designated class, a count of the number of times (U) that state g occurs when P_{gi} falls within the designated class, the sum (ΣP) of the F individual values of P_{gi} within the class, and the sum ($\Sigma\, P(1 - P)$) of the respective products $P_{gi}\,(1 - P_{gi})$ formed from those F individual values of P_{gi} and their complements. The specimen printout shown in table 1 was taken from an analysis of weather data, in which the size of the verification sample was N = 7668 and the predictand state chosen for illustration was g = 4; in the REEP printout, the term "group" is used instead of "state."

Two tests of validity are made on the frequency distributions of P_{gi}. (For justification see reference (12), section 5a, p. 9.) The first of these, denoted by Chi-square$_g$, compares the estimated total number of occurrences of state g with the actual total number of occurrences of state g. This measure, given by the equation Chi-square$_g$ = W_g^2

where $$W_g = \left(\sum_{i=1}^{N} D_{gi} - \sum_{i=1}^{N} P_{gi}\right)/\left[\sum_{i=1}^{N} P_{gi}(1 - P_{gi})\right]^{\frac{1}{2}} \qquad (12)$$

is distributed approximately as chi-square with one degree of freedom. There will be G such statistics. The second test, denoted by Overall chi-square, applies to all G frequency distributions collectively. (The separate values of Chi-square$_g$ cannot be added to obtain an overall test, because they are correlated.) A square matrix $\underline{\rho}$ of order G - 1 is constructed using the following formula for the general element ρ_{ef}:

$$\rho_{ef} = \frac{-S_{ef}}{\sqrt{S_{ee}S_{ff}}} \qquad (e, f = 1, \ldots, G\text{-}1;\ e \neq f) \qquad (13)$$

where

$$S_{ef} = \sum_{i=1}^{N} P_{ei}P_{fi}$$

$$S_{ee} = \sum_{i=1}^{N} P_{ei}(1 - P_{ei})$$

$$S_{ff} = \sum_{i=1}^{N} P_{fi}(1 - P_{fi})$$

TABLE 1

SPECIMEN OF REEP VERIFICATION TABLE

(1) Group (g)	(2) Class	(3) Number of Probabilities (F)	(4) Number of Predictand Occurrences (U)	(5) Sum of Probabilities (ΣP)	(6) Sum of Products ($\Sigma P(1-P)$)
4	$P = 0.0$	12	1	0.	0.
	$0.0<P<0.1$	39	1	2.01	1.88
	$0.1\leq P<0.2$	53	5	8.19	6.89
	$0.2\leq P<0.3$	79	17	20.02	14.88
	$0.3\leq P<0.4$	94	31	32.79	21.27
	$0.4\leq P<0.5$	115	52	51.76	28.37
	$0.5\leq P<0.6$	128	73	70.97	31.52
	$0.6\leq P<0.7$	149	101	97.21	33.67
	$0.7\leq P<0.8$	181	140	137.22	33.05
	$0.8\leq P<0.9$	499	443	428.02	60.46
	$0.9\leq P<1.0$	5019	4951	4917.09	97.29
	$P = 1.0$	1300	1297	1300.00	0.00
Total		7668	7112	7065.27	329.26

Let ρ^{gh} denote the general element of the inverse of the matrix $\underline{\rho}$. The overall validation measure, given by

$$\text{Overall chi-square} = \sum_{g=1}^{G-1} \sum_{h=1}^{G-1} \rho^{gh} W_g W_h \qquad (14)$$

is distributed approximately as chi-square with $G - 1$ degrees of freedom. In the case of just two possible states ($G = 2$) (as with reenlistment or non-reenlistment) this second test is superfluous.

The material shown in table 1 would seem sufficient for still another chi-square test, although its computation is not included in the present program. This is the familiar test of goodness of fit, in which the actual frequency in a class of P_{gi} is given by U and the expected frequency by ΣP. Unfortunately, difficulties arise in the assignment of degrees of freedom. Whereas the expected total frequency of any one state g in the verification sample is not predetermined, the sum of the total expected frequencies over G states is predetermined, being necessarily equal to N, since $\sum_g P_{gi} \equiv 1$. As an approximation, the number of degrees of freedom may be found by subtracting $1/G$ from the number of distinct classes used in computing chi-square for state g.

E. Proof That REEP Probabilities Sum to Unity

Equation (1), defining the REEP regression function, can be written in expanded form as follows:

$$
\begin{aligned}
\hat{D}_1 &= A_{01} + A_{11}X_1 + A_{21}X_2 + \cdots + A_{M1}X_M \\
\hat{D}_2 &= A_{02} + A_{12}X_1 + A_{22}X_2 + \cdots + A_{M2}X_M \\
&\cdot \\
&\cdot \\
&\cdot \\
\hat{D}_G &= A_{0G} + A_{1G}X_1 + A_{2G}X_2 + \cdots + A_{MG}X_M
\end{aligned}
\tag{15}
$$

We shall prove that $A_{01} + A_{02} + \ldots + A_{0G} = 1$ and that $A_{m1} + A_{m2} + \ldots + A_{mG} = 0$ for $m = 1, 2, \ldots, M$, where M is the number of selected predictors. This is sufficient to prove that $\hat{D}_1 + \hat{D}_2 + \ldots + \hat{D}_G \equiv 1$.

The matrix equation for generating the regression coefficients in the g-th equation in (15) is:

$$A_g = C^{-1}X'D_g \tag{16}$$

in which the separate terms are defined as follows:

A_g is a column vector with $M + 1$ elements*, $\{A_{0g} \quad A_{1g} \; \cdots \; A_{Mg}\}$

C is a square matrix of order $M + 1$ consisting of sums, sums of squares and sums of cross-products of the predictor variables,

*Here written in horizontal array for typographical convenience. This expedient will be followed throughout this subsection.

$$C = \begin{bmatrix} N & \Sigma X_1 & \Sigma X_2 & \cdots & \Sigma X_M \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X_2 & \cdots & \Sigma X_1 X_M \\ \vdots & & & & \\ \Sigma X_M & \Sigma X_1 X_M & \Sigma X_2 X_M & \cdots & \Sigma X_M^2 \end{bmatrix} \quad (17)$$

All summations go from 1 to N, where N is the number of cases in the sample.

X' is an M + 1 by N matrix consisting of the individual values of the predictor variables,

$$X' = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ X_{11} & X_{12} & \cdots & X_{1N} \\ X_{21} & X_{22} & \cdots & X_{2N} \\ \vdots & & & \\ X_{M1} & X_{M2} & \cdots & X_{MN} \end{bmatrix} \quad (18)$$

D_g is a column vector with N elements consisting of the individual values of the g-th dummy predictand,

$$D_g = \{D_{1g} \quad D_{2g} \quad \cdots \quad D_{Ng}\} \quad (19)$$

Equation (16) expresses just one set of regression coefficients. The matrix equation for all G sets of coefficients is

$$A = C^{-1}X'D \tag{20}$$

where A is an $M + 1$ by G matrix consisting of the G column vectors $A_1, A_2, \ldots, A_G$. Similarly, D is an N by G matrix consisting of the G column vectors $D_1, D_2, \ldots, D_G$.

Define a column vector e consisting of G elements, each of which is unity: $e = \{1 \;\; 1 \;\; \ldots \;\; 1\}$. Post-multiplying both sides of equation (20) by e gives

$$Ae = C^{-1}X'De \tag{21}$$

Note that Ae gives the sums that we require. That is, the m-th element of Ae is $A_{m1} + A_{m2} + \ldots + A_{mG}$ and m ranges from zero through M.

Consider now the right-hand side of equation (21). De is a column vector with N elements, of which the n-th element is:

$$D_{n1} + D_{n2} + \ldots + D_{nG}$$

This sum is identically equal to unity for all n becuase one and only one of the G states (g) must occur and that D_g takes on the value 1 while the remaining D's take on values of zero.

Next consider $X'De$. This is a column vector with $M+1$ elements:

$$\{N \quad \Sigma X_1 \quad \Sigma X_2 \quad \ldots \quad \Sigma X_M\}$$

This is precisely the first column of the matrix C. Therefore

$$C^{-1}X'De = \{1 \quad 0 \quad 0 \quad \ldots \quad 0\} \tag{22}$$

by the definition of the inverse of a matrix. Thus, referring back to equation (21) we see that

$$\begin{aligned} &A_{01} + A_{02} + \ldots + A_{0G} = 1 \\ &A_{m1} + A_{m2} + \ldots + A_{mG} = 0 \quad (m = 1, 2, \ldots, M) \end{aligned} \tag{23}$$

which was to be proved.

IV. OBJECTIVE DETERMINATION OF CATEGORIES

A. Purpose

The use of dummy variables in predictor screening is exposed to the possibility that the separate dummy components of an ordinary variable might individually be too weak to form an entering wedge in the screening process - with the result that the predictive contribution of that variable will be lost altogether. The point involved here is the important distinction between individual significance and collective significance. On this account, it is advisable to concentrate as much statistical significance as possible in the variables being tested as predictors.

The statistical technique of objective dummying, originated for use within REEP by T. G. Johnson of TRC, has as its purpose the construction of dummy variables well equipped to exhibit significance as predictors in a screening process. This end is achieved by striking a well poised balance between two elements that favor the demonstration of statistical significance but tend to oppose each other. These are the enlargement of sample size within predictor classes and the preservation of predictand differences among predictor classes. The idea, in a nutshell, is to make distinctions among predictor classes when, and only when, there are corresponding predictand differences worth preserving.

B. Procedure

Let the predictand Y be divided into a prescribed number G of operationally significant classes $Y_1, Y_2, \ldots, Y_G$. Let any trial predictor Z be divided as finely as may seem reasonable from the standpoint of physical meaning and observed frequency. Let Z_i and Z_j denote any two categories of Z. By means of the

Kolmogorov-Smirnov test for the discrepancy between two empirical distributions, the array of Y given that Z is in category Z_i is compared with the array of Y given that Z is in category Z_j. The Kolmogorov-Smirnov test is very convenient to use in comparing two distributions, and, like chi-square, is nonparametric. It is a test of significance of the absolute value of the maximum difference between two empirical cumulative relative frequency functions. (See reference (19).)

Two types of comparison can be made:

(1) Ordered comparisons, confined to adjacent categories, or

(2) Unordered comparisons, applicable to arbitrary pairs of categories.

In either case, there is computed a significance measure

$$P(Z_i,Z_j) \quad (0 \leq P(Z_i,Z_j) \leq 1)$$

so defined that a low value of $P(Z_i,Z_j)$ discredits the null hypothesis that the two arrays are samples from the same population, and a high value sustains the null hypothesis. An arbitrary, high level P* of P being set, the program computes P for all pairs admissible under the stipulated kind of comparison (ordered or unordered), determines whether any values of P exceed P*, and if any do exceed P*, the pair of categories of Z yielding the highest value of P is consolidated. This consolidation reduces the number of rows in the contingency table by one. The process is now repeated on the reduced table. When no further rows can be combined at the level P* first chosen, a somewhat lower level of P* is taken, and the process goes forward in the same manner as before. Eventually there is a contingency table in which all rows differ significantly at an assigned nominal level of significance. A nominal level of significance is used rather than an adjusted level, which takes account of the number of comparisons, because the loss of potential predictors is a much more serious risk than that of tentatively admitting a useless variable for further consideration. By the same token, the assigned level is much less strict (say $P^* = .25$ to $.50$) than would be desirable in the usual tests of significance. Typically the cut-off is abrupt: most consolidation takes place at very high levels of P^*, and after a few more consolidations at somewhat lower levels of P^*, the highest remaining values of P fall well below the terminal level for consolidation. At the terminal stage, all variables having at least two significantly different categories are possibly useful predictors. The same process can be applied to the consolidation of compound classes, using either (a) unordered comparisons of all compound classes, or (b) ordered comparisons of categories of one variable for fixed values of the other variables.

V. SIMPLE VS. COMPOUND PREDICTORS

Dummy-variable predictors associated with discrete classes of any individual ordinary variable will be called simple predictors, or univariate dummies. Dummy-variable predictors associated with compound classes, constructed from two or more ordinary variables taken simultaneously, will be called compound predictors, or bivariate dummies in the case of two components, or in general, multivariate dummies. A useful observation for some analytical operations is that multivariate dummies can be constructed by taking the product of corresponding univariate dummies.

Any bounded function $F(Z_1,Z_2,\ldots,Z_r)$ that is defined as the sum of arbitrary bounded functions $F_1(Z_1),F_2(Z_2),\ldots,F_r(Z_r)$ of r separate variables $Z_1,Z_2,\ldots,Z_r$, i.e.,

$$F(Z_1,Z_2,\ldots,Z_r) = F_1(Z_1) + F_2(Z_2) + \ldots + F_r(Z_r) \qquad (24)$$

can be approximated by a linear function of dummy variables associated with the respective Z's, because each component function $F_i(Z_i)$ can be so represented. Hence the use of a linear function of simple predictors implies that the function under consideration is being approximated by a sum of arbitrary functions of the ordinary variables.

An expansion of the type shown in equation (24) implies that the rate of change of F with respect to $Z_i(i=1,2,\ldots,r)$ depends only on Z_i. The simplest type of function that can reflect a dependency of first partial derivatives upon all, or certain sets, of the Z's is one defined as a sum of functions of distinct pairs:

$$F(Z_1,Z_2,\ldots,Z_r) = \sum_{i,j} F_{ij}(Z_i,Z_j) \qquad (25)$$

where the summation extends over appropriate pairs of indices i, j. A function of type shown in equation (25) can be approximated by a linear function of dummy variables associated with bivariate classes of the Z's, because each component function $F_{ij}(Z_i,Z_j)$ can be so represented. Still more complicated functions could be represented with multivariate dummies of higher order, but it is seldom practically feasible to make the attempt. It may also be noted that the sum of two univariate functions such as $F_i(Z_i) + F_j(Z_j)$ could be included as a special case of the general bivariate function $F_{ij}(Z_i,Z_j)$, but it would be inefficient to build up a sum of univariate functions with bivariate dummies. Therefore, when bivariate dummies are employed, it is desirable to make univariate dummies also available as trial predictors. More generally, all dummy variables of lower order should be included in any predictor system employing multivariate dummies.

It is possible to construct examples in which the predictors are exclusively compound terms or in which the significance of simple predictors cannot be detected without including compound predictors. An illustration of the latter situation is the following:

$$\begin{aligned} D &= X_1X_2 + (1 - X_1)(1 - X_2) \\ &\equiv 1 - X_1 - X_2 + 2X_1X_2 \quad . \end{aligned} \tag{26}$$

Here D = 1 when X_1 and X_2 are equal to each other - both predictor-classes being present (1, 1) or both absent (0, 0). D = 0 when X_1 and X_2 are not equal to each other - one predictor-class occuring without the other (0, 1), (1, 0). If all four bivariate predictor-classes (1, 1), (0, 1), (1, 0), (0, 0) are equally likely, neither univariate predictor will show significance until the bivariate dummy X_1X_2 is brought into the regression function.

On the other hand, if the four bivariate classes are not equally likely, the univariate dummies can register significance without the bivariate dummy.

The possibility that interactions might have predictive value should be taken into account, as far as feasible, in drawing up a tentative set of predictors. However, it would be rash to expect radical gains in predictive accuracy from the use of compound predictors. A down-to-earth expectation is that there exists a modest potential of predictability, that might be distributed in appreciably different gradations over a bivariate table. By including bivariate dummies as trial predictors, such gradations could be distinguished.

VI. POPULATION AND VARIABLES USED FOR REENLISTMENT PREDICTION

A. Source of Data

Two sources of data were used to obtain a population to which REEP may be applied for the purpose of obtaining a reenlistment prediction equation. The first source was the active Enlisted Master Tape (EMT) which reflected the characteristics of enlisted personnel serving on active duty in the Navy on 31 July 1963. The EMT is maintained by the Manpower Information Division (Pers 19) of the Bureau of Naval Personnel. The 31 July 1963 active tape (a total of 29 magnetic tapes) was furnished to INS by Pers 19.

The active tape contains a record on each of the 585,000 enlisted men and women on active duty in the Navy. Each record contains about 460 alpha-numeric characters covering about 140 pieces of information needed by the Navy to compile recurring statistical reports.

The second source of data is the set of EMT records of all the enlisted personnel who were lost to the Navy during the time frame under consideration.[*] These records are organized in the same manner as the active EMT records, with one addition -- these "loss tapes" contain the date and type of separation from the Navy. The Manpower Information Division of BuPers furnished a copy of all of the loss tapes relative to the time frame (a total of 11 magnetic tapes).

[*] The time frame of this study covers the period from 1 August 1962 through 31 July 1963, less the month of October 1962. (The records of those enlisted personnel separating from the Navy during October 1962 were not available.)

B. Population

The application of REEP contained in this report concerns itself with the population of USN enlisted men on active duty who were eligible and recommended for reenlistment at some time during the time frame (1 August 1962 through 31 July 1963, less the month of October 1962) and who had served five or fewer years of active Navy duty, had just finished their first enlistment, were in an electronics rating*, and either a) had reenlisted at the end of their term, or b) had reenlisted into the STAR program**, or c) had not reenlisted even though eligible and recommended for reenlistment.

Men whose Current Enlistment Date (CED) fell within the time frame, and who had made continuous service reenlistments (i.e., had reenlisted either immediately or within three months of their Date of Separation (DOS)), form the reenlistee population. Men whose DOS fell within the time frame, and who did not reenlist within three months of their DOS, though they had been eligible and recommended to do so, form the loss population.

This population contained 7,075 men (see table 3 for division between reenlistees and losses, and between developmental and verification samples).

C. Predictor Variables

Those variables used in REEP for prediction of reenlistment were of two types -- univariate and bivariate. In one model, Model A, only the univariates were allowed to be selected for predicting reenlistment action, while in a second model, Model B, the univariates and bivariates were used. In this way we can compare the results of the two models and evaluate how much, if any, improvement in prediction was obtained.

*As defined by DOD Occupational Group 1, Electronics Equipment Repairmen. The ratings included are: SO,TM,FT,MT,ET,DS,AT,ATR, ATN,AX,AQ,TD.

**See the glossary (appendix A) for description of STAR program.

Table 2 lists the 18 univariates used as predictor variables for the REEP analyses. As a starting point in the program, subjective decisions were made concerning how many categories of values each variable will be allowed to assume. These categories were then tested in an objective fashion (see section IV for discussion of the method employed) using the developmental sample[1] to determine whether any further consolidation of categories could be made. This approach yielded 61 different categories of values corresponding to the 18 variables. (The 61 categories that were considered in Model A as a result of the objective determination are listed together with other pertinent data in table 4).

TABLE 2

REEP UNIVARIATE PREDICTOR VARIABLES[2]

1. Years of Education
2. Education Difference
3. Education Ratio
4. General Classification Test
5. Arithmetic Test
6. Mechanical Test
7. GCT + ARI
8. Electronics Technician Selection Test
9. Race
10. Age At Initial Enlistment
11. Recruiting Area
12. State Unemployment Rate[3]
13. Migration Index[4]
14. Median State Income[3]
15. % With Income in State[3]
16. Median State Income[4]
17. Months of Previous Military Duty
18. Birthplace/Residence

[1] See section II for explanation of the "developmental" and "verification" samples.

[2] See appendix A for description of these variables.

[3] For Whites

[4] For Non-Whites

For the bivariate analysis, eight of the 18 univariates were chosen and all possible pairs of these eight were then considered. The eight variables chosen are:

1. Years of Education
2. GCT + ARI
3. ETST Score
4. Race
5. Age at Initial Enlistment
6. Recruiting Area
7. Median State Income (for Whites)
8. Education Difference.

In addition to the possible pairs formed in this way, one additional bivariate was considered, viz., that formed by the simultaneous consideration of "age at initial enlistment" and "length of previous military duty."

The consideration of all of these pairs of variables led to 331 bivariate subcategories. It was found that a) many of these bivariate subcategories contained very few men (fewer than 20) of the developmental sample and b) certain subcategories formed by some pairs of variables had similar reenlistment rates (the small differences being possibly due to chance). Hence, the bivariate subcategories were tested in the (unordered) objective grouping program to determine whether any grouping of subcategories could be made. As a result it was found that some subcategories may be combined and others may be eliminated from consideration altogether (the latter effect taking place when it is found that for a certain category of one variable there are only insignificant differences in reenlistment rate for all categories of the second variable, thus implying that no bivariate effect exists there). In this manner, it was found that only 153 bivariate subcategories need be considered as predictor variables. Thus, in Model B, there were 214 predictor variables considered -- 61 univariates and 153 bivariates.

The 153 bivariate subcategories are shown in tables 5 through 29 (which also show the reenlistment rate and percent of population in each subcategory for the developmental sample).

D. Predictand

We have just discussed the variables that will be used for prediction and to some extent how the choice was made. For the REEP application being described, the predictand (i.e., that variable which we are attempting to predict) is reenlistment action. Based on the predictors chosen by REEP as being significantly related to reenlistment, REEP assigns to each man a probability of reenlistment.

VII. RESULTS

A. General Results

The overall reenlistment rates which applied to the first term electronics population used in the REEP analyses are shown in table 3 for the developmental and verification samples. The close correspondence of the rates between the two samples acted as supporting evidence of the randomness of the split of the population in forming the developmental (90%) and verification (10%) samples.

TABLE 3

GENERAL STATISTICS PERTAINING TO DEVELOPMENTAL AND VERIFICATION SAMPLES

	Developmental	Verification	Total
Reenlistees	1647	179	1826
Losses	4725	524	5249
Total	6372	703	7075
Reenlistment Rate	.258	.255	.258
% Sample	90%	10%	100%

The reenlistment rates and population distribution are shown for the developmental sample

i) in table 4 as a function of the univariate predictors,

and ii) in tables 5 through 29 as a function of the bivariate predictors.

It is noted that in table 4 the "% of Sample" column does not add up to exactly 100% for most of the variables. (The same situation exists in tables 5 through 29.) This is attributed to one or more of the following reasons: a) There is some slight error introduced when the numbers are rounded-off for entry into the table. b) There was no information available on that variable for some men. c) The variable simply did not apply to certain men.

It is further noted that in the case of the variable "Race" the percentages add up to 101.1%. This is due to the fact that Negroes are considered both separately (Category 2) and in the Non-Caucasian group (Category 3).

Tables 5 through 29 show the bivariate categories used in Model B as well as data pertaining to each category. The categories are numbered sequentially starting with #62 (there were 61 univariate categories as shown in table 4). In addition to the variable number, each box shows the reenlistment rate and the percentage of the population that fell in that bivariate category.

TABLE 4

REEP UNIVARIATE PREDICTOR VARIABLES:
CATEGORY VALUES AND OTHER DATA

Number	Variable	Category	Values	Developmental Sample	
				Reenl. Rate	% of Sample
1	Years of	1	< 12	.311	12.5
2	Education	2	12	.241	76.1
3		3	> 12	.317	11.4
4	Education	1	< -2.00	.419	0.7
5	Difference	2	-2.00 to 0.99	.274	32.1
6		3	1.00 to 1.99	.228	50.3
7		4	≥ 2.00	.315	16.4
8	Education	1	< .75	.478	0.4
9	Ratio	2	.75 to 0.99	.285	14.9
10		3	1.00 to 1.249	.242	71.7
11		4	≥ 1.25	.318	12.5
12	General	1	≤ 44	.194	1.0
13	Classification	2	45 - 50	.291	3.2
14	Test	3	51 - 61	.219	40.3
15	(GCT)	4	62 - 66	.261	31.5
16		5	≥ 67	.315	23.7
17	Arithmetic	1	0 - 51	.267	8.1
18	Test	2	52 - 56	.208	16.8
19	(ARI)	3	57 - 62	.244	40.9
20		4	≥ 63	.295	33.9
21	Mechanical	1	≤ 47	.192	12.7
22	Test	2	48 - 52	.224	19.4
23		3	53 - 61	.257	40.3
24		4	≥ 62	.311	27.3
25	GCT + ARI	1	≤ 89	.194	0.6
26		2	90 - 99	.305	1.9
27		3	100 - 124	.221	54.8
28		4	125 - 134	.281	32.6
29		5	≥ 135	.373	9.9
30	Electronics	1	≤ 50	.260	3.1
31	Technician	2	51 - 63	.219	41.8
32	Selection	3	64 - 66	.248	19.0
33	Test	4	≥ 67	.296	31.1

TABLE 4 (Cont.)

				Developmental Sample	
Number	Variable	Category	Values	Reenl. Rate	% of Sample
34	Race	1	Caucasian	.256	98.7
35		2	Negro	.500	1.1
36		3	Non-Caucasian	.459	1.3
37	Age At Initial Enlistment	1	< 18.5	.232	52.5
38		2	18.5 - 19	.266	34.5
39		3	20 - 22	.322	11.8
40		4	≥ 23	.595	1.2
41	Recruiting Area	1	1,2,4,5,6	.237	62.4
42		2	3,7	.321	19.0
43		3	8	.268	18.0
44	State (1) Unemployment Rate*	1	≤ 3.8	.246	30.8
45		2	3.9 - 4.6	.307	15.6
46		3	≥ 4.7	.246	51.8
47	Migration (2) Index**	1	≤ -17	.364	0.2
48		2	> -17	.500	0.6
49	Median State (3) Income*	1	≤$3800	.301	25.4
50		2	> 3800	.240	72.9
51	% With (4) Income In State*	1	≤ 89.0	.300	12.0
52		2	89.9 - 92.5	.245	79.6
53		3	≥ 92.9	.301	6.6
54	Median State (3) Income**	1	≤$3500	.556	0.8
55		2	> 3500	.308	0.4
56	Months of Previous Military Duty	1	None (5)	.251	93.5
57		2	13 - 24	.315	2.6
58		3	25 - 60	.377	2.9
59		4	≥ 61	.697	0.5
60	Birthplace/ Residence	1	Same	.245	70.8
61		2	Different	.298	25.3

* For Whites

**For Non-Whites

(1) See table A-6

(2) See table A-3

(3) See table A-2

(4) See table A-4

(5) Includes 0-12 months.

The purpose of presenting tables 4 through 29 in this section is not to perform analyses on the data themselves, but as auxiliary information pertaining to the description of the population and variables used. It is certainly legitimate to perform analyses of the reenlistment rates and population distributions shown in these tables but such analyses are of contingency count type and are not included in the design and intent of REEP. Rather, it is proper to use these tables as reference material when, for example, the significant predictors are finally selected through REEP. One can then refer back to the appropriate table(s) to learn more about the behavior of the population with respect to the selected predictors.

TABLE 5

REEP BIVARIATES AND ASSOCIATED DATA: YEARS OF EDUCATION VS. GCT + ARI

		Years of Education		
		< 12	12	> 12
GCT+ ARI	≤ 89	#62 .250 (0.3%)	#68 .211 (47.8%)	
	90-99	#63 .351 (0.9%)		
	100-124	#64 .284 (8.2%)		
	125-134	#65 .378 (2.7%)	#67 .256 (25.2%)	#69 .370 (14.2%)
	≥ 135	#66 .318 (0.3%)		

TABLE 6

REEP BIVARIATES AND ASSOCIATED DATA: YEARS OF EDUCATION VS. ETST SCORE

		Years of Education	
		12	> 12
ETST Score	≤ 50	#70 .270 (1.8%)	#74 .225 (2.8%)
	51-63	#71 .201 (32.4%)	
	64-66	#72 .235 (15.1%)	#75 .309 (2.1%)
	≥ 67	#73 .278 (23.3%)	#76 .355 (6.2%)

TABLE 7

REEP BIVARIATES AND ASSOCIATED DATA: YEARS OF EDUCATION VS. RACE

		Years of Education	
		< 12	12
Race	Caucasian	#77 .308 (12.3%)	#78 .238 (75.3%)
	Non-Caucasian	#79 .508 (1.0%)	

TABLE 8

REEP BIVARIATES AND ASSOCIATED DATA: YEARS OF EDUCATION VS. AGE AT INITIAL ENLISTMENT

		Years of Education		
		< 12	12	> 12
Age At Initial Enlistment	< 18.5	#80 .285(10.2%)	#82 .219(42.1%)	#84 .312(16.7%)
	18.5, 19	#81 .431(2.3%)	#83 .251(27.8%)	
	20-22			
	≥ 23		#85 .587 (1.0%)	

TABLE 9

REEP BIVARIATES AND ASSOCIATED DATA:
YEARS OF EDUCATION VS. RECRUITING AREA

		Years of Education		
		< 12	12	> 12
Recruiting Area	1,2,4,5,6	#86 .290 (6.1%)	#88 .223 (50.1%)	#90 .299 (6.2%)
	3,7	#87 .393 (3.2%)	#89 .286 (12.8%)	#91 .397 (3.0%)

TABLE 10

REEP BIVARIATES AND ASSOCIATED DATA:
YEARS OF EDUCATION VS. MEDIAN STATE INCOME (WHITES)

		Years of Education		
		< 12	12	> 12
Median State Income (Whites)	≤ $3800	#92 .374 (3.8%)	#94 .275 (18.1%)	#96 .359 (3.5%)
	> $3800	#93 .277 (8.4%)	#95 .226 (57.0%)	#97 .301 (7.5%)

TABLE 11

REEP BIVARIATES AND ASSOCIATED DATA: YEARS OF EDUCATION VS. EDUCATION DIFFERENCE

Education Difference	Years of Education < 12	12	> 12
< -2.00	#98 .419(0.7%)		
-2.00 to 0.99	#103 .311(28.0%)	#99 .260 (20.4%)	#101 .189(0.8%)
1.00 to 1.99		#100 .221(47.6%)	#102 .374(1.9%)
≥ 2.00			

TABLE 12

REEP BIVARIATES AND ASSOCIATED DATA: GCT + ARI VS. ETST SCORE

ETST Score	GCT + ARI ≤ 99	100-124	125-134	≥ 135
≤ 50	#105 .252 (22.0%)			
51-63		#104 .208 (41.5%)		
64-66			#106 .287 (24.2%)	
≥ 67				#107 .382 (7.3%)

TABLE 13

REEP BIVARIATES AND ASSOCIATED DATA: GCT + ARI VS. RACE

		GCT + ARI				
		≤ 89	90-99	100-124	125-134	≥ 135
Race	Caucasian	#108 .152 (0.5%)	#109 .273 (1.7%)	#110 .217 (53.9%)	#113 .282 (32.4%)	#114 .373 (10.0%)
	Negro			#111 .500 (0.8%)		
	Non-Caucasian *			#112 .482 (0.9%)		

*Includes Negro, American Indian, Malayan, and Mongolian

TABLE 14

REEP BIVARIATES AND ASSOCIATED DATA: GCT + ARI VS. AGE AT INITIAL ENLISTMENT

		GCT + ARI		
		≤ 124	125-134	≥ 135
Age At Initial Enlistment	<18.5	#115 .215 (50.6%)	#117 .246 (17.6%)	#120 .351 (7.8%)
	18.5,19		#118 .302 (10.8%)	
	20-22	#116 .253 (5.9%)	#119 .356 (3.9%)	#121 .458 (1.9%)

TABLE 15

REEP BIVARIATES AND ASSOCIATED DATA: GCT + ARI VS. RECRUITING AREA

		GCT + ARI		
		≤ 99	100-124	125-134
Recruiting Area	1,2,4,5,6	#123 .201 (34.8%)		#126 .253 (21.2%)
	3,7	#122 .475 (0.6%)	#124 .279 (11.8%)	#127 .372 (5.0%)
	8	#125 .225 (9.7%)		#128 .302 (6.2%)

TABLE 16

REEP BIVARIATES AND ASSOCIATED DATA: GCT + ARI VS. MEDIAN STATE INCOME (WHITES)

		GCT + ARI		
		≤ 99	100-124	125-134
Median State Income (Whites)	≤$3800	#129 .404(0.7%)	#130 .260(15.1%)	#132 .344(7.3%)
	>$3800	#131 .199 (40.1%)		#133 .264(25.0%)

TABLE 17

REEP BIVARIATES AND ASSOCIATED DATA: GCT + ARI VS. EDUCATION DIFFERENCE

		GCT + ARI		
		≤ 99	100-124	125-134
Education Difference	≤ 0.99	#134 .310(1.3%)	#136 .244(18.8%)	#139 .305(10.1%)
	1.00 to 1.99	#135 .250(1.1%)	#137 .197(27.8%)	#140 .240(16.5%)
	≥ 2.00		#138 .253(7.9%)	#141 .354(5.8%)

TABLE 18

REEP BIVARIATES AND ASSOCIATED DATA: ETST SCORE VS. RACE

		ETST Score			
		≤ 50	51-63	64-66	≥ 67
Race	Caucasian	#142 .247 (3.0%)	#143 .215 (41.3%)	#146 .247 (18.7%)	#148 .297 (31.0%)
	Negro	#144 .606 (0.5%)		#147 .417 (0.2%)	
	Non-Caucasian*	#145 .605 (0.6%)		#149 .242 (0.5%)	

*Includes Negro, American Indian, Malayan, and Mongolian

TABLE 19

REEP BIVARIATES AND ASSOCIATED DATA: ETST SCORE VS. AGE AT INITIAL ENLISTMENT

Age At Initial Enlistment	ETST Score			
	≤ 50	51-63	64-66	≥ 67
< 18.5	#153 .210 (35.5%)			#154
18.5,19	#150 .358(0.8%)	#151 .232(14.1%)	.275(32.6%)	
20-22	#152 .239 (4.1%)		#155 .364(7.1%)	

TABLE 20

REEP BIVARIATES AND ASSOCIATED DATA: ETST SCORE VS. RECRUITING AREA

Recruiting Area	ETST Score		
	51-63	64-66	≥ 67
1,2,4,5,6	#157 .210 (37.4%)		#159 .274(20.3%)
3,7	#156 .279(8.8%)	#160 .336 (14.1%)	
8	#158 .217 (10.7%)		

TABLE 21

REEP BIVARIATES AND ASSOCIATED DATA: ETST SCORE VS. MEDIAN STATE INCOME (WHITES)

		ETST Score			
		≤ 50	51-63	64-66	≥ 67
Median State Income (Whites)	≤$3800	#161 .346(0.8%)	#162 .256(11.9%)	#165 .327 (11.1%)	
	>$3800	#163 .199 (31.4%)		#164 .226(14.1%)	#166 .286(23.9%)

TABLE 22

REEP BIVARIATES AND ASSOCIATED DATA: ETST SCORE VS. EDUCATION DIFFERENCE

		ETST Score			
		≤ 50	51-63	64-66	≥ 67
Education Difference	≤ 0.99	#171 .246 (22.0%)			#172 .323(8.6%)
	1.00 to 1.99	#167 .250(1.1%)	#169 .190(21.1%)	#173 .241 (25.9%)	
	≥ 2.00	#168 .375(0.4%)	#170 .245(5.8%)	#174 .348 (9.6%)	

TABLE 23

REEP BIVARIATES AND ASSOCIATED DATA: AGE AT INITIAL ENLISTMENT VS. RECRUITING AREA

		Age at Initial Enlistment		
		< 18.5	18.5,19	20-22
Recruiting Area	1,2,4,5,6	#175 .210(33.3%)	#178 .245(21.7%)	#179 .311(7.0%)
	3,7	#176 .283(8.9%)	#180 .345 (9.7%)	
	8	#177 .254(10.2%)	#181 .277 (7.6%)	

TABLE 24

REEP BIVARIATES AND ASSOCIATED DATA: AGE AT INITIAL ENLISTMENT VS. MEDIAN STATE INCOME (WHITES)

		Age At Initial Enlistment			
		≤ 18.5	18.5,19	20-22	≥ 23
Median State Income (Whites)	≤$3800	#182 .260(12.1%)	#185 .325 (12.8%)		#187 .724(0.4%)
	>$3800	#183 .219(39.8%)	#184 .246(24.7%)	#186 .308(7.7%)	#188 .513(0.6%)

TABLE 25

REEP BIVARIATES AND ASSOCIATED DATA:
AGE AT INITIAL ENLISTMENT VS. EDUCATION DIFFERENCE

		Age At Initial Enlistment	
		< 18.5	18.5,19
Education Difference	≤ 0.99	#191 .271 (30.7%)	
	1.00 to 1.99	#189 .198 (27.1%)	#192 .240(18.2%)
	≥ 2.00	#190 .268 (4.3%)	#193 .321(6.4%)

TABLE 26

REEP BIVARIATES AND ASSOCIATED DATA:
RECRUITING AREA VS. MEDIAN STATE INCOME (WHITES)

		Recruiting Area
		1,2,4,5,6
Median State Income (Whites)	≤$3800	#194 .267 (8.4%)
	>$3800	#195 .229 (53.3%)

TABLE 27

REEP BIVARIATES AND ASSOCIATED DATA:
RECRUITING AREA VS. EDUCATION DIFFERENCE

		Recruiting Area		
		1,2,4,5,6	3,7	8
Education Difference	≤ 0.99	#196 .259(11.5%)	#199 .348(5.2%)	#200 .267 (16.1%)
	1.00 to 1.99	#197 .220(41.2%)	#201 .265 (9.1%)	
	≥ 2.00	#198 .283(9.8%)	#202 .363 (6.6%)	

TABLE 28

REEP BIVARIATES AND ASSOCIATED DATA:
MEDIAN STATE INCOME (WHITES) VS. EDUCATION DIFFERENCE

		Median State Income (Whites)	
		≤ $3800	> $3800
Education Difference	-2.00 to 0.99	#203 .341 (5.1%)	#205 .260(26.8%)
	1.00 to 1.99	#204 .272 (11.2%)	#206 .213(38.8%)

TABLE 29

REEP BIVARIATES AND ASSOCIATED DATA:
MONTHS OF PREVIOUS MILITARY DUTY VS.
AGE AT INITIAL ENLISTMENT

		Months of Previous Military Duty			
		None	13-24	25-60	≥ 61
Age At Initial Enlistment	< 18.5	#207 .232(51.2%)	#211 .283(0.8%)	#213 .343(2.6%)	
	18.5,19	#208 .260(31.9%)	#212 .330(1.8%)		
	20-22	#209 .316(10.0%)			
	≥ 23	#210 .400(0.4%)		#214 .694(0.8%)	

B. REEP Model A

The variables used in this application of REEP were of two types -- univariate and bivariate. In one model, Model A, only univariates were allowed to be selected for predicting reenlistment action, while in a second model, Model B, both univariates and bivariates were made available. In this way we can compare the results of the two models and evaulate how much, if any, improvement in prediction is obtained by including bivariates as well as univariates.

This section discusses the results obtained using Model A. Section VII-C discusses results of Model B, and section VII-D compares the results of the two models.

Of 61 dummy variables under consideration as possible predictors in Model A, seven were selected as significant by the screening procedure. Accordingly, the REEP regression function for estimating reenlistment rate is of the form:

$$\hat{D}_1 = B_0 + B_1X_1 + \ldots + B_7X_7 \qquad (27)$$

The corresponding function $\hat{D}_2$ for estimating non-reenlistment rate is simply the complement of $\hat{D}_1$ (i.e., $\hat{D}_2 \equiv 1 - \hat{D}_1$) and need not be considered further, except to mention the fact that the Brier P-score*, being summed over two groups (G=2), will be exactly twice the mean-square error for either group (reenlistment or non-reenlistment) by itself. The selected predictors (designated by X's) and values of the regression coefficients (B's) are given in table 30. The notation X_1 represents the most significant predictor, X_2 the second most significant predictor, or more

*See section III-D for a description of the Brier P-score.

precisely, the most significant adjunct to X_1, and so on in order of subscript, so that X_7 represents the seventh most significant predictor, or more precisely, the most significant adjunct to X_1 through X_6. The F-ratios for each selected predictor were well above the respective critical values of F. For example, the F-ratio for X_7 was 17.8 as compared with the critical level of 11.6. However, the F-ratio for the eighth most significant predictor fell below the critical level; hence, only seven predictors were accepted as being significant.

TABLE 30

TERMS IN $\hat{D}_1$ FOR MODEL A

Predictor Symbol*		Additive Constant	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Regression Coefficients	Symbol	B_0	B_1	B_2	B_3	B_4	B_5	B_6	B_7
	Value	.272	-.063	.292	.075	.059	.281	-.047	.082

*Exact identification of the predictors selected as significant for Model A is available to qualified users upon request to the Director for Long Range Studies (Op-911).

To use the data in table 30 for deriving a reenlistment rate for any given individual is a relatively simple matter because of the use of dummy variables. The coefficients of the characteristics that pertain to the individual are added to the baseline, the "additive constant." For example, if none of the seven selected categories pertain to an individual, then his predicted reenlistment rate is simply .272 (the additive constant alone), which is a little above average. If categories 1, 2, and 4 pertain, then his predicted probability of reenlistment is .560 (calculated as .272 - .063 + .292 + .059), which is high.

The Brier P-score for the developmental sample (N=6372) was 0.3703; that for the verification sample (N=703) was 0.3707 . Hence, overall predictive performance, as measured by the Brier P-score, was very nearly the same on the verification sample as on the developmental sample, thus lending credence to the estimated P-score. Of course, if the population itself were to undergo basic changes in the relationships among the variables, the present regression function should not be expected to apply.

Tabular summaries of actual numbers of reenlistments versus expected numbers based on REEP estimates of reenlistment probability, in both the developmental and verification samples, are given in table 31.

TABLE 31

ACTUAL REENLISTMENTS VS. ESTIMATED RATES - MODEL A

Class of Predicted Reenlistment Probability (P)		Developmental Sample			Verification Sample		
		No. Elig. (F)	Number Reenlisted		No. Elig. (F)	Number Reenlisted	
			Actual (U)	Expected* (ΣP)		Actual (U)	Expected (ΣP)
Low	$P=0.0$	0	0	0.00	0	0	0.00
	$0.0<P<0.1$	0	0	0.00	0	0	0.00
	$0.1\leq P<0.2$	1140	199	185.10	128	24	20.78
Avg.	$0.2\leq P<0.3$	3738	866	910.50	404	100	98.37
High	$0.3\leq P<0.4$	1117	403	374.13	133	34	44.40
	$0.4\leq P<0.5$	281	125	119.89	30	15	12.77
	$0.5\leq P<0.6$	59	35	32.35	7	6	3.87
	$0.6\leq P<0.7$	26	14	16.63	1	0	0.69
	$0.7\leq P<0.8$	8	3	5.84	0	0	0.00
	$0.8\leq P<0.9$	3	2	2.56	0	0	0.00
	$0.9\leq P<1.0$	0	0	0.00	0	0	0.00
	$P=1.0$	0	0	0.00	0	0	0.00
Total		6372	1647	1647.00	703	179	180.88

*Expected on the basis of Model A predictions.

The value of Chi-square (obtained by setting g=1 in equation (12), section III-D) for the verification sample* was 0.03 (with one degree of freedom). This small value of chi-square indicates very close agreement between the actual total number of reenlistments and the expected total number of reenlistments.

As mentioned in section III-D, a more stringent test of agreement between actual and expected numbers of reenlistments, by classes of P, is afforded by the chi-square test of goodness of fit. If the population values of reenlistment rate were known exactly as a function of the predictors, the predictand (here, actual number of reenlistments) would exhibit a random distribution (of binomial form, with the parameter n being equal to the number of eligibles) about the expected number. In testing a random sample from such a population, the value of chi-square should lie in the neighborhood of its median value, for best vindication of the hypothesized probability function.

Numerical quantities used in the calculation of chi-square for this test of goodness of fit are displayed in table 32. There are five classes of P, and it will be assumed that there are 4½ degrees of freedom.** By linear interpolation between 4 and 5 degrees of freedom in standard tables, the value of chi-square at the 50% level of significance (i.e., the median) is estimated to be 3.85, and the value at the 25% level of significance is estimated to be 6.01. The value of chi-square

*By definition of the REEP regression function, the corresponding measure for the developmental sample would be identically zero, if all estimates were admissible. In practice, the actual computed value usually rounds to zero, for the developmental sample.

**See last paragraph of section III-D.

TABLE 32

CHI-SQUARE TEST OF GOODNESS OF FIT FOR VERIFICATION SAMPLE - MODEL A

Class of P	Number Reenlisted		(Act-Exp)	$\frac{(\text{Act-Exp})^2}{\text{Exp}}$
	Actual	Expected*		
$P < 0.2$	24	20.78	3.22	0.4990
$0.2 \leq P < 0.3$	100	98.37	1.63	0.0270
$0.3 \leq P < 0.4$	34	44.40	-10.40	2.4360
$0.4 \leq P < 0.5$	15	12.77	2.23	0.3894
$P \geq 0.5$	6	4.56	1.44	0.4547
Totals	179	180.88	-1.88	3.8061

*Expected on the basis of Model A predictions.

calculated from the verification sample of Model A (entry in lower right-hand corner of table 32) is 3.81. Therefore, valid estimates of reenlistment rates are obtained using the hypothesized probability function of the form shown in equation (27) with coefficients as shown in table 30.

Once the validity of the probability estimates has been established, there remains the issue of degree of resolution: how well can poorer risks be distinguished from better risks? Probability estimates can be valid without having much resolving power; for instance, they can be clustered so closely around the average that no consequential information on departure from average can be obtained from them. Evidence bearing on departure from average can be found in table 31. For the verification sample of Model A, this evidence can be summarized by stating that (1) it is possible to identify a group, comprising about 18% of the eligibles, for which the reenlistment rate (.188) is significantly lower than average, and (2) it is possible to identify another group, comprising about 24% of the eligibles, for which the reenlistment rate (.322) is significantly higher than average. Obviously, it is possible also to identify the remaining group, comprising about 58% of the eligibles, for which the reenlistment rate (.248) is about average.

C. REEP Model B

In Model B, both univariate and bivariate predictors were made available for selection. The results on Model B parallel closely those on Model A, and a similar style of description will be used.

Of 214 dummy variables (61 univariates and 153 bivariates) under consideration as possible predictors in Model B, it turned out, again, that seven were selected as significant by the screening procedure. The REEP regression function for estimating reenlistment rate thus reduced to the same general form as in Model A, namely

$$\hat{D}_1 = C_0 + C_1Y_1 + \dots + C_7Y_7 \qquad (28)$$

This time, however, the selected predictors were different from before, and of course so were the regression coefficients. The selected predictors (designated by Y's) and values of the coefficients (C's) are shown in table 33. As in Model A, the subscripts indicate the rank order of selection. The notation Y_1 represents the most significant predictor, Y_2 the second most significant predictor, or more precisely, the most significant adjunct to Y_1, and so on in order of subscript, with Y_7 representing the seventh most significant predictor, or more precisely, the most significant adjunct to Y_1 through Y_6. The F-ratios for each selected predictor were well above the respective critical values of F. For example, the F-ratio for Y_7 was 17.7 as compared with the critical level of 14.3. However, the F-ratio for the eighth most significant predictor fell below the critical level; hence, only seven predictors were accepted.

TABLE 33

TERMS IN $\hat{D}_1$ FOR MODEL B

Predictor Symbol*		Additive Constant	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7
Regression Coefficients	Symbol	C_0	C_1	C_2	C_3	C_4	C_5	C_6	C_7
	Value	.257	-.068	.393	-.053	.096	.338	.056	.071

*Exact identification of predictors selected as significant for Model B is available to qualified users upon request to the Director for Long Range Studies (Op-911).

The Brier P-score for the developmental sample of Model B (N = 6372) was 0.3687; that for the verification sample (N = 703) was 0.3694. The indicated superiority of Model B over Model A, shown by the slightly (but statistically significantly) lower P-score of the former in the developmental sample, was borne out in the verification sample.

Tabular summaries of actual numbers of reenlistments versus expected numbers based on REEP estimates of reenlistment probability, in both the developmental and verification samples, are given in table 34.

TABLE 34

ACTUAL REENLISTMENTS VS. ESTIMATED RATES - MODEL B

Class of Predicted Reenlistment Probability (P)		Developmental Sample			Verification Sample		
		No. Elig. (F)	Number Reenlisted		No. Elig. (F)	Number Reenlisted	
			Actual (U)	Expected* (ΣP)		Actual (U)	Expected* (ΣP)
Low	$P=0.0$	0	0	0.00	0	0	0.00
	$0.0<P<0.1$	0	0	0.00	0	0	0.00
	$0.1\leq P<0.2$	1770	319	303.14	187	36	31.53
Avg.	$0.2\leq P<0.3$	2787	643	686.05	313	79	76.80
High	$0.3\leq P<0.4$	1400	492	467.12	163	47	54.44
	$0.4\leq P<0.5$	329	137	134.65	33	10	13.51
	$0.5\leq P<0.6$	36	20	21.17	3	3	1.79
	$0.6\leq P<0.7$	23	20	14.87	3	3	1.97
	$0.7\leq P<0.8$	23	14	16.58	1	1	0.72
	$0.8\leq P<0.9$	3	1	2.41	0	0	0.00
	$0.9\leq P<1.0$	1	1	0.99	0	0	0.00
	$P=1.0$	0	0	0.00	0	0	0.00
Total		6372	1647	1646.98	703	179	180.76

*Expected on the basis of Model B predictions.

The value of Chi-square (obtained by setting g=1 in equation (12), section III-D) for the verification sample was 0.02, indicating excellent agreement between the actual total number of reenlistments and the expected total number of reenlistments.

As before, a more stringent test of validity was carried out by calculating chi-square for measuring goodness of fit (see table 35). The value obtained was 4.04, which is very close to the median value, and hence vindicates the validity of the REEP probability estimates obtained using the hypothesized function of the form shown in equation (28) with coefficients as shown in table 33.

TABLE 35

CHI-SQUARE TEST OF GOODNESS OF FIT
FOR VERIFICATION SAMPLE - MODEL B

Class of P	Number Reenlisted		(Act-Exp)	$\frac{(\text{Act-Exp})^2}{\text{Exp}}$
	Actual	Expected*		
$P < 0.2$	36	31.53	4.47	0.6337
$0.2 \leq P < 0.3$	79	76.80	2.20	0.0630
$0.3 \leq P < 0.4$	47	54.44	-7.44	1.0168
$0.4 \leq P < 0.5$	10	13.51	-3.51	0.9119
$P \geq 0.5$	7	4.48	2.52	1.4175
Total	179	180.76	-1.76	4.0429

*Expected on the basis of Model B predictions.

Degree of resolution was examined in the manner described under Model A (section VII-B). Evidence bearing on departure from average is presented in table 34. Referring to results on the verification sample of Model B, it can be seen that (1) it is possible to identify a group, comprising nearly 27% of the eligibles, for which the reenlistment rate (.193) is significantly lower than average, and (2) it is possible to identify another group, comprising about 29% of the eligibles, for which the reenlistment rate (.315) is significantly higher than average. It is therefore possible also to identify the remaining group, comprising about 44% of the eligibles, for which the reenlistment rate (.252) is about average.

D. Comparison of Models

It has been shown that both models yield valid estimates of reenlistment probability, but the evidence on degree of resolution suggests that Model B has greater capacity for sorting out departures from average. This apparent difference between models can be tested for significance by applying the chi-square test for homogeneity of two empirical distributions. This test (see table 36) was made on the verification samples, using a simplified formula for chi-square (a special case of the Brandt-Snedecor formula*) applicable when comparing two samples of equal size. Strictly, such a test requires independent samples. However, since dependence caused by using records on the same men would only tend to make the two models appear more nearly alike than they would with two different groups of men, a large value of chi-square will imply greater significance than it would

*Reference (20), section 9.10, p. 205.

with independent samples. The value found for chi-square was 25.85, and with 4 degrees of freedom this is highly significant (the critical value for the 0.001 level is 18.47). Therefore, Model B makes significantly sharper distinctions than does Model A.

TABLE 36

CHI-SQUARE TEST FOR HOMOGENEITY OF DISTRIBUTIONS OF P IN MODELS A AND B

Class of P	Frequency in Model A (a)	Frequency in Model B (b)	(b+a)	(b-a)	$\frac{(b-a)^2}{b+a}$
$P < 0.2$	128	187	315	59	11.05
$0.2 \leq P < 0.3$	404	313	717	-91	11.55
$0.3 \leq P < 0.4$	133	163	296	30	3.04
$0.4 \leq P < 0.5$	30	33	63	3	0.14
$P \geq 0.5$	8	7	15	-1	0.07
Total	703	703	1406	0	25.85

On comparing the predictors selected under the two models, it was found that six of the seven selected predictors are more or less closely related between the two models.

REFERENCES

(1) Bancroft, T. A., "On Biases in Estimation Due to Preliminary Tests of Significance," The Annals of Mathematical Statistics, Vol. 15, pp. 190-204, 1944.

(2) Burington and May, "Handbook of Probability and Statistics," Handbook Publishing, Inc., Sandusky, Ohio, 1953.

(3) Cox, D. R., "Two Further Applications of a Model for Binary Regression," Biometrika, Vol. 45, pp. 562-565, 1958.

(4) Cox, D. R., "The Regression Analysis of Binary Sequences," Journal of the Royal Statistical Society, Series B, Vol. 20, pp. 215-242, 1958.

(5) DOD Occupational Table, Enlisted, Oct 1963.

(6) Enger, I., Russo, J. A., Jr., and Sorenson, E. L. "A Statistical Approach to 2-7 hr. Prediction of Ceiling and Visibility," Vol. I Tech. Rpt. 7411-118, Contract Cwb-10704, The Travelers Research Center, Inc., Hartford, Conn., 1964.

(7) Gumbel, E. J., "Statistics of Extremes," (Second Printing), Columbia University Press, New York, 1960.

(8) Lund, I. A., "Estimating the Probability of a Future Event from Dichotomously Classified Predictors," Bulletin of the American Meteorological Society, Vol. 36, No. 7, pp. 325-328, 1955.

(9) Lund, I. A., "Some Application of the Method of Least Squares to Estimating the Probability of a Future Event," GRD Research Notes, No. 51, Geophysics Research Directorate, Air Force Cambridge Research Laboratories, Bedford, Mass., 13 pp., 1961.

(10) Maxwell, A. E., "Analyzing Qualitative Data," Methuen and Co., Ltd., London, 1961

(11) Miller, R. G., "Regression Estimation of Event Probabilities," Tech. Rpt. 7411-121, Contract Cwb-10704, The Travelers Research Center, Inc., Hartford, Conn., 1964.

(12) Miller, R. G., "Statistical Prediction by Discriminant Analysis," Meteorological Monographs, Vol. 4, No. 25, American Meteorological Society, Boston, Mass., 54 pp., 1962.

(13) NavPers 15642, "Naval Manpower Information System," Parts I and II, BuPers, 1959.

(14) NavPers 15791A, "Bureau of Naval Personnel Manual."

(15) NavPers 15949A, "Manual of the Active Duty Enlisted Master Magnetic Tape Record," BuPers, Oct. 1962.

(16) Oosterhoff, J., "On the Selection of Independent Variables in a Regression Equation," p. 18, Report S 319 (VP 23) Stichting Mathematisch Centrum, Amsterdam, 1963.

(17) Singer, A., et al., "Multivariate Study of Enlisted Retention," INS Doc. No. U-65-10 789. Institute of Naval Studies, Cambridge, Mass. For Official Use Only. October, 1965.

(18) Singer, A., Bryan, J. G., et al., "Multivariate Study of Enlisted Retention -- Phase II," INS Document No. U-65-11 064. Institute of Naval Studies, Cambridge, Mass. For Official Use Only. October, 1965.

(19) Smirnov, N., "Tables for Estimating the Goodness of Fit of Empirical Distributions," The Annals of Mathematical Statistics, Vol. 19, pp. 279-281, 1948.

(20) Snedecor, G. W., "Statistical Methods," (Fourth Printing), The Iowa State College Press, Ames, Iowa, 1950.

(21) "United States Census of Population, 1960, United States Summary, General Social and Economic Characteristics," U. S. Department of Commerce, Bureau of the Census, U. S. Government Printing Office.

(22) U. S. Census of Population: 1960, Vol. I, Characteristics of the Population, Parts 1-57, Chapter C. General Social and Economic Characteristics, Washington, U. S. Department of Commerce, Bureau of Census.

(23) Warner, S. L., "Multivariate Regression of Dummy Variates Under Normality Assumptions," Journal of the American Statistical Association; Vol. 58, No. 304, pp. 1054-1063, 1963.

APPENDIX A

GLOSSARY

Arithmetic (ARI) Test Score -- Indicates an individual's ability to use numbers and apply mathematical reasoning in practical problems. It is one of the four tests that make up the Navy's Basic Test Battery.

Birthplace/Residence -- Indicates whether the state of birth of an individual is the same as or different from his state of residence at time of entry into the Navy.

Contingency Table -- Refers to a table of frequencies of classifications based on discrete values of two or more variables.

Correlation -- Denotes degree of dependence of one variable on another.

Developmental Sample -- Refers to that portion of the population used in the REEP program for the purpose of analysis in choosing the significant predictor variables and assigning their coefficients.

DOD Occupational Group -- Groupings of enlisted men by armed service occupational skills along lines defined by the Department of Defense. (reference (5)).

Dummy Variable -- Refers to an index used to denote the occurrence or non-occurrence of a designated event. The index has only 2 admissible values, 0 or 1. When the designated event occurs, the index value is 1; when the event does not occur, the index value is 0.

Education Difference -- Reflects the individual's perceived relative educational level as measured by the difference between the number of years of education of men 25 years or older of the same race from the same state of residence. The measure is given by the formula:

EDUCATION DIFFERENCE =

$$\frac{\text{(Individual's Educ.) - (Median Educ. of Same Race in Same State)}}{\text{Standard Deviation of Median Educ. of Same Race}},$$

where the denominator is equal to 1.095 for whites and 1.568 for non-whites. The median education for race and state is given in table A-1.

This variable was used only for men from the 48 contiguous states.

Education Ratio -- Refers to the ratio of the number of years of an individual's education to the median number of years of education of men 25 years or older of the same race from the same state of residence. The median education for race and state is given in table A-1.

This variable was used only for men from the 48 contiguous states.

TABLE A-1

MEDIAN YEARS OF EDUCATION FOR MALES 25 YEARS OR OLDER BY RACE AND STATE*

	Race	
State	White	Non-White
Alabama	10.0	5.8
Arizona	11.4	6.8
Arkansas	9.0	5.7
Calif.	12.1	10.2
Colorado	11.9	11.1
Conn.	10.8	8.8
Delaware	11.3	8.0
Florida	11.4	6.4
Georgia	10.0	5.3
Idaho	11.2	9.7
Illinois	10.6	8.9
Indiana	10.7	8.7
Iowa	10.4	9.0
Kansas	11.3	9.1
Kentucky	8.6	7.8
La.	10.2	5.4
Maine	10.5	10.9
Maryland	10.8	7.7
Mass.	11.3	10.0
Michigan	10.6	8.7
Minn.	9.9	9.6
Miss.	10.6	5.1
Missouri	9.4	8.5
Montana	10.8	8.7
Nebraska	10.9	9.3
Nevada	12.1	8.6
N. H.	10.5	11.0
N. J.	10.8	8.6
N. M.	11.3	7.1
N. Y.	10.8	9.1
N. C.	9.2	6.1
N. D.	8.9	8.2
Ohio	10.8	8.8
Oklahoma	10.3	8.4
Oregon	11.2	9.5
Pa.	10.1	8.7
R. I.	10.0	9.4
S. C.	9.8	5.1
S. D.	9.1	8.5
Tenn.	8.8	6.9
Texas	10.5	7.6
Utah	12.2	10.1
Vermont	10.0	11.1
Va.	10.2	6.5
Wash.	11.8	10.2
W. Va.	8.7	7.8
Wisc.	9.9	8.8
Wyoming	11.7	9.4

*Data taken from table 47 ofreference (22).

Electronics Technician Selection Test (ETST) Score -- Indicates an individual's abilities that are specifically related to successful completion of electronics training, and his understanding of and/or familiarity with mathematics, science (physics), shop practice, electricity, and radio. It is used as an aid in selecting personnel for training as electronics technicians.

Electronics Men -- Refers to those men whose Navy rating places them in the DOD occupational group classified as Electronic Equipment Repairmen. The Navy electronics ratings are: SO, TM, FT, MT, ET, DS, AT, ATR, ATN, AX, AQ, TD. The service rating CTM is included in the DOD definition of electronics occupations, but could not be so included for this study since the information on the service rating of individuals in the general CT rating was not available.

Eligible for Reenlistment -- All reenlistees are assumed to have been eligible for reenlistment. The fact that a non-reenlistee had or had not been eligible for reenlistment is reflected by a special code on the man's loss tape record.

Enlisted Master Tapes (EMT) -- The magnetic tapes which contain the personnel records of the enlisted Navy.

GCT + ARI -- Refers to the sum of the individual's scores on the General Classification Test (GCT) and the Arithmetic (ARI) Test.

General Classification Test (GCT) Score -- Indicates an individual's ability to understand words and relationships between words and ideas, thus indirectly measuring reasoning ability. It is one of the four tests that make up the Navy's Basic Test Battery.

Least Squares -- Refers to the method of determining adjustable constants in a mathematical function in such a way that the sum of squares of the differences between the individual values of the dependent variable and the function being used to approximate them is made as small as possible.

Level of Significance -- Refers to the probability that an observed value of a variable will differ from the theoretical value by the observed amount if the conditions of some assumed hypothesis were correct.

Loss -- Refers to a person who separates from the Navy.

Loss Tapes -- Refers to the Enlisted Master Tapes containing data on those who separated from the Navy.

Mechanical (MECH) Test Score -- Indicates an individual's aptitude for mechanical work, mechanical and electronic knowledge, and ability to understand mechanical principles. It is one of the four tests that make up the Navy's Basic Test Battery.

Median -- Refers to that value of a variable which is greater than or equal to half of the observed values of the variable, and less than or equal to the other half.

Median Income -- Identifies the median income of men 14 years or older, by race, in a man's state of residence. Table A-2 gives these values.

This variable was used only for men from the 48 contiguous states.

TABLE A-2

MEDIAN INCOME FOR MALES 14 YEARS OR OLDER BY RACE AND STATE*

State	Race	
	White	Non-White
Alabama	$3367	$1417
Arizona	4262	1845
Arkansas	2486	993
Calif.	5109	3515
Colorado	4228	3163
Conn.	5033	3516
Delaware	4879	2421
Florida	3743	2073
Georgia	3374	1489
Idaho	3866	1987
Illinois	5056	3613
Indiana	4456	3448
Iowa	3708	3141
Kansas	3968	2636
Kentucky	2928	1764
La.	4001	1565
Maine	3275	1970
Maryland	4875	2756
Mass.	4422	2984
Michigan	4984	3728
Minn.	4012	2616
Miss.	2757	890
Missouri	3851	2570
Montana	3993	1461
Nebraska	$3497	$2882
Nevada	5076	3184
N. H.	3845	2492
N. J.	5172	3341
N. M.	4101	2009
N. Y.	4798	3307
N. C.	3035	1286
N. D.	3134	1416
Ohio	4903	3433
Oklahoma	3446	1613
Oregon	4470	3019
Pa.	4348	3216
R. I.	3848	2503
S. C.	3195	1135
S. D.	3043	964
Tenn.	2932	1598
Texas	3728	1917
Utah	4558	2739
Vermont	3320	2029
Va.	3734	1906
Wash.	4689	2989
W. Va.	3470	2097
Wisc.	4417	3631
Wyoming	4435	1977

*Data taken from table 67 in reference (22).

Migration Index -- Reflects a weighted measure of the difference between white and non-white migration out of the individual's state of residence, and was derived in the following manner. The net migration of whites was subtracted from the net migration of non-whites for each state (this yielded a rough index of movement of non-whites due primarily to differences in economic opportunities for whites and non-whites). The difference was them multiplied by the ratio of the percent of the state's non-white population in 1960 to the percent of the United States' non-white population for the same year. (See table A-3; basic data drawn from reference (21).)

Model A -- Refers to the predictive system used in this study, which employed only univariate components that were combined through a linear function of dummy variables to yield an estimate, obtained through REEP, of the probability of reenlistment of eligible first term electronics men.

Model B -- Refers to the predictive system used in this study, which employed both univariate and bivariate components that were combined in a linear function of dummy variables to yield an estimate, obtained through REEP, of the probability of reenlistment of eligible first term electronics men.

Multiple Correlation -- The correlation between the actual values of the predictand and corresponding estimated values given by a regression function using two or more predictors. The coefficient of multiple correlation lies in the range 0-1.

Net Effect -- Refers to the residual predictive component ascribed to a given predictor after allowance has been made for the simultaneous effects of the other predictors being used.

TABLE A-3

MIGRATION INDEX*

State	Index
Alabama	41.87
Arizona	54.96
Arkansas	30.46
Calif.	-17.15
Colorado	-10.87
Conn.	-22.99
Delaware	7.78
Florida	83.63
Georgia	47.04
Idaho	-1.85
Illinois	-26.99
Indiana	-12.85
Iowa	-2.06
Kansas	-3.70
Kentucky	-3.84
La.	35.84
Maine	-3.95
Maryland	7.73
Mass.	-7.59
Michigan	-22.53
Minn.	-1.81
Miss.	86.60
Missouri	-10.95
Montana	2.33
Nebraska	-6.10
Nevada	-6.60
N. H.	-5.90
N. J.	-18.50
N. M.	7.88
N. Y.	-23.36
N. C.	33.78
N. D.	0.26
Ohio	-15.71
Oklahoma	2.91
Oregon	-4.04
Pa.	-11.70
R. I.	-3.48
S. C.	80.00
S. D.	1.83
Tenn.	4.19
Texas	5.29
Utah	-1.10
Vermont	-0.79
Va.	23.29
Wash.	-7.84
W. Va.	5.67
Wisc.	-14.87
Wyoming	2.19

*This index was applied to non-whites only.

Null Hypothesis -- Refers to an initial hypothesis which is to be tested.

Objective Dummying -- Refers to the objective procedure for dividing a predictor variable into classes with the goal of maximizing the statistical significance of the associated dummy variables as predictors of a stated dependent variable.

Ordinary Variable -- Refers to a variable in the usual sense of the term. The qualifier "ordinary" is used synonymously with "initial" or "raw" to indicate a natural or less extensively processed state.

Percent with Income -- Identifies by race the percent of men 14 years or older who have some source of income in the man's state of residence. Table A-4 gives these values.

This variable was used only for men from the 48 contiguous states.

Predictand -- Refers to that variable which one is attempting to predict, for example, reenlistment. (This term is used synonymously with "dependent variable.")

Predictor Variables -- Refers to those variables to be used for purposes of predicting an event. (This term is used synonymously with "independent variables.")

Previous Military Duty -- Indicates the amount of military service an individual had served in one of the other armed services before entering the Navy. This amount is reflected by the difference between two dates on the Enlisted Master Tape.

TABLE A-4

PERCENT OF MALES 14 YEARS OR OLDER WITH INCOME BY RACE AND STATE*

State	Race	
	White	Non-White
Alabama	87.6	80.6
Arizona	90.7	74.9
Arkansas	88.4	85.2
Calif.	92.1	87.8
Colorado	92.4	89.7
Conn.	91.6	87.1
Delaware	91.4	87.0
Florida	90.8	86.2
Georgia	88.8	83.5
Idaho	92.9	86.8
Illinois	91.7	85.3
Indiana	91.2	85.0
Iowa	91.3	87.2
Kansas	92.5	88.2
Kentucky	85.8	84.3
La.	88.1	81.8
Maine	91.9	95.2
Maryland	90.8	84.3
Mass.	91.4	88.9
Michigan	90.8	81.8
Minn.	91.3	85.8
Miss.	87.6	81.7
Missouri	90.7	85.0
Montana	92.0	88.1
Nebraska	91.6	86.8
Nevada	93.2	85.4
N. H.	92.4	91.5
N. J.	90.9	85.9
N. M.	89.0	70.7
N. Y.	90.1	85.8
N. C.	88.3	81.7
N. D.	90.0	75.6
Ohio	91.3	85.3
Oklahoma	90.7	82.5
Oregon	93.4	88.8
Pa.	89.9	84.2
R. I.	91.2	86.2
S. C.	88.6	78.7
S. D.	88.9	77.4
Tenn.	87.2	83.4
Texas	90.4	86.2
Utah	92.3	85.5
Vermont	91.4	88.5
Va.	90.0	83.1
Wash.	93.5	90.1
W. Va.	84.5	79.9
Wisc.	91.7	86.7
Wyoming	93.2	92.8

*Data taken from table 67 in reference (22).

Rating -- Identifies the occupation (rating or apprenticeship) in which an individual is serving.

Recommended for Reenlistment -- In order to reenlist, an individual must not only be eligible for reenlistment but must also be recommended for reenlistment by his commanding officer. All reenlistees are assumed to have been recommended for reenlistment. The fact that a loss who was eligible for reenlistment had or had not been recommended for reenlistment is reflected on his loss tape record. (Recommendation for reenlistment is not applicable to losses who were not eligible for reenlistment.)

Recruiting Area -- Refers to one of eight officially prescribed geographical areas defining regions for which recruiting quotas are set. (See figure A-1 and table A-5)

Reenlistment Action -- An individual who has either reenlisted or separated from the Navy in a given period of time is said to have taken a reenlistment action in that time period.

Reenlistment Rate -- If R denotes the number of reenlistees, and L denotes the number of losses that were eligible and recommended for reenlistment (so that R+L is the number of people that were eligible and recommended for reenlistment), then the reenlistment rate is defined as the ratio R/(R+L):

$$\text{Reenlistment Rate} = \frac{\text{Number of Reenlistees}}{\text{Number Eligible and Recommended for Reenlistment}}$$

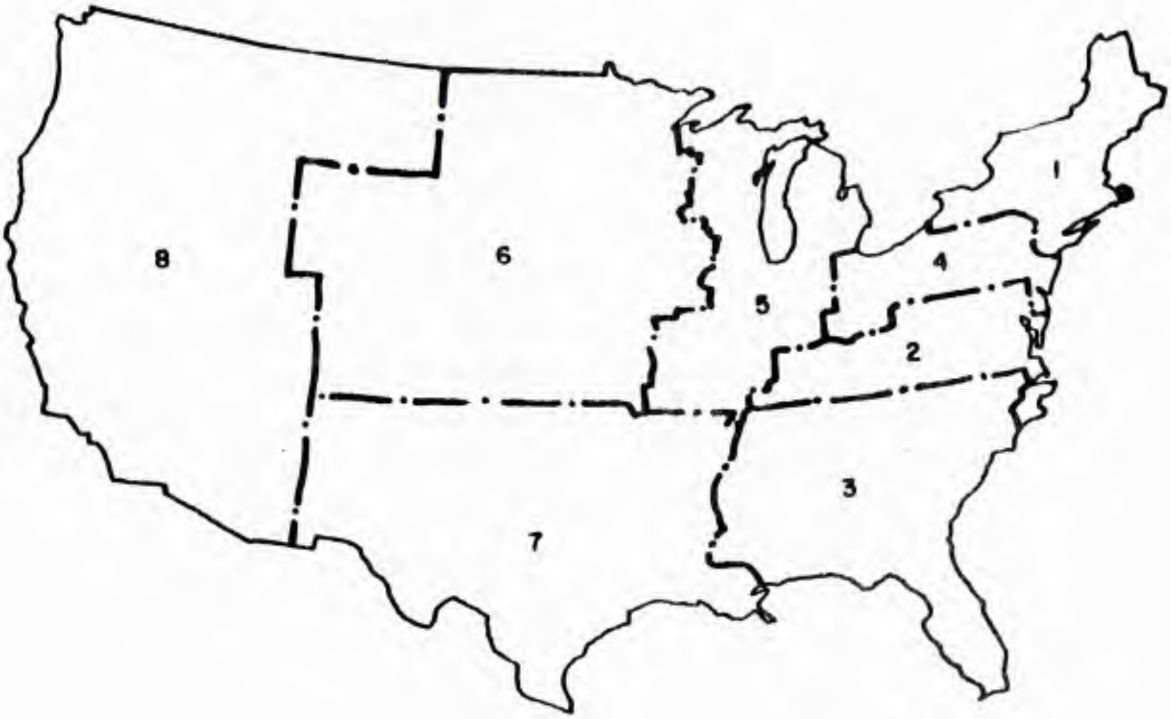

FIG. A-1: MAP OF RECRUITING AREAS

TABLE A-5

KEY TO RECRUITING AREAS*

Area	States
1	Maine, N. H., Vt., Mass., Conn., R. I., N. Y., N. J.
2	Ky., Md., Va., W. Va.
3	Ga., Fla., Ala., Miss., Tenn., N. C., S. C.
4	Del., Pa., Ohio
5	Mo., Ind., Ill., Mich., Wis.
6	Minn., Iowa, Nebr., Kansas, N. Dak., S. Dak., Colo., Wyo.
7	Ark., La., Tex., Okla., N. Mex.
8	Idaho, Mont., Wash., Oreg., Nev., Utah, Ariz., Calif.

*Although the southern part of the state of New Jersey officially fell in Area 4, the state was placed in Area 1, the area containing the major part of the state's population. Similarly, North Carolina was placed in Area 3 although the northeastern corner of the state officially fell in Area 2.

REEP -- Regression Estimation of Event Probabilities

Regression -- Refers to the statistical dependence of one variable upon stated other variables. A regression function is the mathematical expression of the mean value of one (dependent) variable -- the predictand -- as a function of other (independent) variables -- the predictors.

Screening Procedure -- Refers to the statistical process of screening predictor variables and ranking them preferentially in order of their incremental contributions to the predictand.

STAR Program -- The Selective Training and Retention Program instituted by the Navy in August 1960 as an incentive for reenlistment in the hopes of raising the reenlistment rate in certain critical ratings.

State of Residence -- Identifies the state or United States possession in which an individual officially resided at the time of initial entry into the naval service. Residence in the Republic of the Philippines or other foreign country is also identified.

Stepwise Regression -- Refers to the sequential selection of variables to be used as predictors in a regression function.

TRC -- The Travelers Research Center, Inc. (Hartford, Connecticut).

Unemployment Rate -- Identifies the rate of unemployment, by race, for men 14 years or older in the state of residence of an individual. Table A-6 gives these values.

This variable was used only for men from the 48 contiguous states.

Verification Sample -- Refers to that portion of the population set aside in the REEP program for use in testing the predictive ability of the variables and coefficients which were chosen based on the developmental sample.

Years of Education -- Indicates the total number of years (grades) of formal schooling completed by the individual at the time of his initial entry into the Navy.

TABLE A-6

PERCENT UNEMPLOYED FOR MALES 14 YEARS OR OLDER BY RACE AND STATE*

	Race	
State	White	Non-White
Alabama	4.7	8.4
Arizona	4.6	14.4
Arkansas	4.9	8.6
Calif.	5.5	10.1
Colorado	3.8	6.7
Conn.	3.7	8.7
Delaware	3.9	10.1
Florida	4.6	6.5
Georgia	3.3	5.7
Idaho	5.4	8.3
Illinois	3.7	11.0
Indiana	3.7	7.7
Iowa	3.2	10.8
Kansas	3.4	9.3
Kentucky	6.1	9.1
La.	4.9	10.3
Maine	6.2	17.9
Maryland	3.8	9.4
Mass.	4.2	8.4
Michigan	5.9	16.9
Minn.	5.3	15.4
Miss.	4.3	6.3
Missouri	3.8	8.7
Montana	6.1	26.9

	Race	
State	White	Non-White
Nebraska	2.9	8.0
Nevada	5.5	9.9
N. H.	3.9	6.7
N. J.	3.5	8.9
N. M.	5.4	16.0
N. Y.	4.7	7.7
N. C.	3.6	7.4
N. D.	5.7	31.1
Ohio	4.9	12.6
Oklahoma	4.0	10.1
Oregon	5.9	11.1
Pa.	6.1	12.2
R. I.	4.6	10.3
S. C.	2.9	5.4
S. D.	3.6	27.8
Tenn.	4.9	6.6
Texas	4.0	7.3
Utah	3.9	6.1
Vermont	4.4	15.0**
Va.	3.5	7.1
Wash.	6.0	13.4
W. Va.	9.2	13.5
Wisc.	3.7	11.6
Wyoming	4.9	10.9

*Data taken from table 53 of reference (22).

**Estimated value. No value given in data source.

reenlistment
prediction
navy
regression
analysis
REEP